Towards a General Theory of Infinite-Width Limits of Neural Classifiers: Supplementary Material

Eugene A. Golikov [1]

References

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

A. Experimental details

We perform our experiments on a feed-forward net with $H+1$ hidden layers with no biases. We learn our network as a binary classifier on a subset of CIFAR2 dataset (which is a dataset of first two classes of CIFAR10) of size 1000. We train our network for 50 epochs to minimize the binary cross-entropy loss and report the final cross-entropy loss on a full test set (of size 2000). We repeat our experiments for 5 random seeds and report means and standard deviations on our plots. We experiment with other setups (i.e. using a mini-batch gradient estimation instead of the exact one, using a larger train dataset, using more training steps, learning a multi-class classification problem) in SM I. All experiments were conducted on a single NVIDIA GeForce GTX 1080 Ti GPU using pytorch framework (Paszke et al., 2017). Our code is available online: https://github.com/deepmipt/research/tree/master/Infinite_Width_Limits_of_Neural_Classifiers.

Although our analysis assumes initializing variables with samples from a gaussian, nothing changes if we sample $\sigma\xi$ for ξ being any symmetric random variable with a distribution independent on hyperparameters.

In our experiments we took a network of width $d^* = 2^7 = 128$ and apply a Kaiming uniform initialization scheme (He et al., 2015) to its layers; we call this network a reference network. Consider a network with a single hidden layer first. According to the Kaiming initialization scheme, initial weights should have zero mean and standard deviations $\sigma_a^* \propto (d^*)^{-1/2}$ and $\sigma_w^* \propto d_0^{-1/2}$, where d_0 is the input dimension which we do not modify. For this network we take (unscaled!) learning rates $\eta_a^* = \eta_w^* = 0.02$ for the gradient descent training and $\eta_a^* = \eta_w^* = 0.0002$ and $\beta = 0.99$ for the RMSProp training. After that, we scale the initial weights and the learning rates with width d according to a specific scaling:

$$\sigma = \sigma^* \left(\frac{d}{d^*}\right)^{q_\sigma}, \quad \hat{\eta}_{a/w} = \hat{\eta}_{a/w}^* \left(\frac{d}{d^*}\right)^{\tilde{q}_{a/w}}.$$

Since $\sigma = \sigma_a \sigma_w$ and since we apply the (leaky) ReLU non-linearity, we can take

$$\sigma_a = \sigma_a^* \left(\frac{d}{d^*}\right)^{q_\sigma}, \quad \sigma_w = \sigma_w^*.$$

Since for GD we have $\hat{\eta}_{a/w} = \eta_{a/w}/\sigma_{a/w}^2$, then

$$\eta_a = \eta_a^* \left(\frac{\sigma_a}{\sigma_a^*}\right)^2 \left(\frac{d}{d^*}\right)^{\tilde{q}_a} = \eta_a^* \left(\frac{d}{d^*}\right)^{\tilde{q}_a + 2q_\sigma},$$

$$\eta_w = \eta_w^* \left(\frac{\sigma_w}{\sigma_w^*}\right)^2 \left(\frac{d}{d^*}\right)^{\tilde{q}_w} = \eta_w^* \left(\frac{d}{d^*}\right)^{\tilde{q}_w}.$$

Similar holds for the multi-layer case. In this case since $\sigma = (\sigma_a \sigma_{v^H} \dots \sigma_{v^1} \sigma_w)^{1/(1+H)}$, we can take

$$\sigma_{a/v^1/\dots/v^H} = \sigma_{a/v^1/\dots/v^H}^* \left(\frac{d}{d^*}\right)^{q_\sigma}, \quad \sigma_w = \sigma_w^*.$$

[1]Neural Networks and Deep Learning lab., Moscow Institute of Physics and Technology, Moscow, Russia. Correspondence to: Eugene A. Golikov <golikov.ea@mipt.ru>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

B. Dynamics of the limit model for the NTK scaling

First consider a continuous-time gradient descent for a one-hidden layer network in a general form:

$$\dot{\theta}_d^{(t)} = -\hat{\eta}\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f(\mathbf{x};\,\theta_d^{(t)})}\frac{\partial f(\mathbf{x};\,\theta_d^{(t)})}{\partial \theta_d},$$

where $\theta_d^{(t)} = \{(\hat{a}_r^{(t)}, \hat{\mathbf{w}}_r^{(t)})\}_{r=1}^d$ is a sequence of d weights $(\hat{a}, \hat{\mathbf{w}})$ associated with each neuron at a time-step t.

$$\dot{f}(\mathbf{x}';\,\theta_d^{(t)}) = \left(\frac{\partial f(\mathbf{x}';\,\theta_d^{(t)})}{\partial \theta_d}\right)^T \dot{\theta}_d^{(t)} =$$

$$= -\hat{\eta}\mathbb{E}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=\langle\ldots\rangle}\left(\frac{\partial f(\mathbf{x}';\,\theta_d^{(t)})}{\partial \theta_d}\right)^T \frac{\partial f(\mathbf{x};\,\theta_d^{(t)})}{\partial \theta_d} =$$

$$= -\hat{\eta}\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f(\mathbf{x};\,\theta_d^{(t)})}\Theta_d(\mathbf{x}',\mathbf{x};\,\theta_d^{(t)}).$$

Assume the model is scaled as $d^{-1/2}$:

$$f(\mathbf{x};\,\theta_d^{(t)}) = d^{-1/2}\sum_{r=1}^d \hat{a}_r^{(t)}\phi(\hat{\mathbf{w}}_r^{(t),T}\mathbf{x}).$$

Then a neural tangent kernel is written as follows:

$$\Theta_d(\mathbf{x}',\mathbf{x};\,\theta_d^{(t)}) = d^{-1}\sum_{r=1}^d\Big(\phi(\hat{\mathbf{w}}_r^{(t),T}\mathbf{x})\phi(\hat{\mathbf{w}}_r^{(t),T}\mathbf{x}') + \\ +\hat{a}_r^2\phi'(\hat{\mathbf{w}}_r^{(t),T}\mathbf{x})\phi'(\hat{\mathbf{w}}_r^{(t),T}\mathbf{x}')\mathbf{x}^T\mathbf{x}'\Big).$$

If moreover $\hat{\eta} = \text{const}$, then for a fixed t independent of d $\hat{a}^{(t)} \to \hat{a}^{(0)}$ and $\hat{\mathbf{w}}^{(t)} \to \hat{\mathbf{w}}^{(0)}$. Hence due to the Law of Large Numbers $\Theta_d(\mathbf{x}',\mathbf{x};\,\theta_d^{(t)}) \to \Theta_\infty(\mathbf{x}',\mathbf{x})$, where

$$\Theta_\infty(\mathbf{x}',\mathbf{x}) = \mathbb{E}_{\,(\hat{a},\hat{\mathbf{w}})\sim\mathcal{N}(0,I_{1+d_0})}\Big(\phi(\hat{\mathbf{w}}^T\mathbf{x})\phi(\hat{\mathbf{w}}^T\mathbf{x}') + \\ +\hat{a}^2\phi'(\hat{\mathbf{w}}^T\mathbf{x})\phi'(\hat{\mathbf{w}}^T\mathbf{x}')\mathbf{x}^T\mathbf{x}'\Big).$$

In the case of the discrete-time dynamics we have similarly:

$$\theta_d^{(k+1)} = \theta_d^{(k)} - \hat{\eta}\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f(\mathbf{x};\,\theta_d^{(k)})}\frac{\partial f(\mathbf{x};\,\theta_d^{(k)})}{\partial \theta_d}.$$

A classical result of calculus states that there exists a $\xi_d^{(k)} \in [0,1]^{(d_0+1)d}$ such that following holds:

$$f(\mathbf{x}';\,\theta_d^{(k+1)}) - f(\mathbf{x}';\,\theta_d^{(k)}) = \\ = \left(\frac{\partial f(\mathbf{x}';\,\tilde{\theta}_d^{(k)})}{\partial \theta_d}\right)^T(\theta_d^{(k+1)} - \theta_d^{(k)}) =$$

$$= -\hat{\eta}\mathbb{E}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=\langle\ldots\rangle}\left(\frac{\partial f(\mathbf{x}';\,\tilde{\theta}_d^{(k)})}{\partial \theta_d}\right)^T\frac{\partial f(\mathbf{x};\,\theta_d^{(k)})}{\partial \theta_d} =$$

$$= -\hat{\eta}\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f(\mathbf{x};\,\theta_d^{(t)})}\Theta_d(\mathbf{x}',\mathbf{x};\,\theta_d^{(k)},\tilde{\theta}_d^{(k)}).$$

where $\tilde{\theta}_d^{(k)} = \theta_d^{(k+1)} \odot \xi_d^{(k)} + \theta_d^{(k)} \odot (1 - \xi_d^{(k)})$, and we have abused notation by redefining Θ_d. Nevertheless, in this case $\Theta_d(\mathbf{x}',\mathbf{x};\,\theta_d^{(k)},\tilde{\theta}_d^{(k)})$ still converges to $\Theta_\infty(\mathbf{x}',\mathbf{x})$ defined above for the same reasons as above.

C. Validation of the power-law asumptions

In Section 3 we have introduced power-law assumptions for weight increments and for terms of the model decomposition:

$$|\delta\hat{a}_r^{(k)}| \propto d^{q_a^{(k)}}, \quad \|\delta\hat{\mathbf{w}}_r^{(k)}\| \propto d^{q_w^{(k)}}; \tag{1}$$

$$f_{d,\emptyset}^{(k)}(\mathbf{x}) \propto d^{q_{f,\emptyset}^{(k)}},\ f_{d,a/w}^{(k)}(\mathbf{x}) \propto d^{q_{f,a/w}^{(k)}},\ f_{d,aw}^{(k)}(\mathbf{x}) \propto d^{q_{f,aw}^{(k)}}. \tag{2}$$

After that, we have derived corresponding exponents for two cases: $q_{a/w}^{(1)} = q_\sigma + \tilde{q}_{a/w} < 0$ and $q_{a/w}^{(1)} = q_\sigma + \tilde{q}_{a/w} = 0$, where q_σ is an exponent for σ and $\tilde{q}_{a/w}$ are exponents for learning rates:

$$\sigma \propto d^{q_\sigma}, \quad \hat{\eta}_{a/w} \propto d^{\tilde{q}_{a/w}}.$$

In order to have a non-vanishing non-diverging limit model $f_\infty^{(k)}$ that does not coincide with its initialization $f_\infty^{(0)}$, we have derived a set of conditions: see Section 3. For the first case these conditions were the following:

$$q_\sigma \in (-1, -1/2],$$

$$q_{a/w}^{(1)} \le -1 - q_\sigma, \quad \max(q_a^{(1)}, q_w^{(1)}) = -1 - q_\sigma,$$

while for the second case they were:

$$q_\sigma = -1, \quad q_{a/w}^{(1)} = 0.$$

The MF scaling is exactly the second case, while the NTK scaling corresponds to the first case: $q_\sigma = q_a^{(1)} = q_w^{(1)} = -1/2$. We have refered a family of scalings $q_\sigma \in (-1, -1/2)$ and $q_a^{(1)} = q_w^{(1)} = -1 - q_\sigma$ as "intermediate".

As we have also derived in Section 3, for both cases $q_{a/w}^{(k)} = q_{a/w}^{(1)}\ \forall k \ge 1$.

Here we validate power-law assumptions (1) as well as derived values for corresponding exponents for the three special cases noted above: MF, NTK and intermediate scalings, see Figure 1. We train a one hidden layer network with the gradient descent for 50 epochs; see SM A for further details. We take norms of final learned weight increments and average them over hidden neurons:

$$\text{av.}\,|\delta\hat{a}^{(k)}| = \frac{1}{d}\sum_{r=1}^{d}|\delta\hat{a}_r^{(k)}|,$$

$$\text{av.}\,\|\delta\hat{\mathbf{w}}^{(k)}\|_2 = \frac{1}{d}\sum_{r=1}^{d}\|\delta\hat{\mathbf{w}}_r^{(k)}\|_2.$$

We then plot these values as functions of width d.

As one can see on left and center plots, weight increments as functions of width are very well fitted with power-laws for both input and output layers. Right plot matches numerical estimates for corresponding exponents $q_a^{(k)}$ and $q_w^{(k)}$ with their theoretical values (denoted by red ticks). Here we notice a reasonable coincidence between them.

In order to validate a power-law assumption for model decomposition terms (2), we compute the variance with respect to the data distribution for each decomposition term. The reason to consider variances instead of decomposition terms themselves is that these terms are functions of $\mathbf{x}$. If we just fix a random $\mathbf{x}$, then the numerical estimate for, say, $f_{d,a}^{(k)}(\mathbf{x})$ can be noisy. Hence it is better to plot some statistics of these terms with respect to the data, hoping that this statistics will be more robust, which is true e.g. for expectation. However, since we consider a binary classification problem with balanced classes, we are likely to have $\mathbb{E}_{\mathbf{x}} f_d^{(k)}(\mathbf{x}) \approx 0$. Because of this, we are afraid to have all of the decomposition terms to be approximately zeros in expectation. For this reason, we consider a variance instead of the expectation. Note that $f_d^{(k)} \propto d^{q_f^{(k)}}$ implies $\mathbb{V}\text{ar}_{\,x} f_d^{(k)} \propto d^{2q_f^{(k)}}$.

As we see in Figure 2, variances of all of the model decomposition terms are fitted with power-laws well. The only exception is $\mathbb{V}\text{ar}_{\,x} f_{d,\emptyset}^{(k)}(\mathbf{x})$ for the mean-field scaling: see the solid curve on the left plot. Nevertheless, this term converges to a constant for large d, which indicates that our analysis becomes valid at least in the limit of large d. Note that we have also matched numerical estimates of corresponding exponents with their theoretical values in Figure 1 of the main text.

D. Derivation of $\varkappa$-terms in a one hidden layer case

For the sake of completeness, we copy all necessary definitions from Section 3 here. A gradient descent step is defined as follows:

$$\Delta\delta\hat{a}_r^{(k)} = -\hat{\eta}_a\sigma\mathbb{E}\,\nabla_f^{(k)}\ell\;\phi((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x}),$$

$$\Delta\delta\hat{\mathbf{w}}_r^{(k)} = -\hat{\eta}_w\sigma\mathbb{E}\,\nabla_f^{(k)}\ell\;(\hat{a}_r^{(0)} + \delta\hat{a}_r^{(k)})\phi'(\ldots)\mathbf{x}, \quad (3)$$

$$\delta\hat{a}_r^{(0)} = 0,\;\delta\hat{\mathbf{w}}_r^{(0)} = 0,\;\hat{a}_r^{(0)} \sim \mathcal{N}(0,1),\;\hat{\mathbf{w}}_r^{(0)} \sim \mathcal{N}(0, I);$$

$$f_d^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^{d}(\hat{a}_r^{(0)} + \delta\hat{a}_r^{(k)})\phi((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x}).$$

We assume:

$$\sigma \propto d^{q_\sigma}, \quad \hat{\eta}_{a/w} \propto d^{\tilde{q}_{a/w}}.$$

$$|\delta\hat{a}_r^{(k)}| \propto d^{q_a^{(k)}}, \quad \|\delta\hat{\mathbf{w}}_r^{(k)}\| \propto d^{q_w^{(k)}}. \quad (4)$$

Assuming our model $f_d^{(k)}$ does not diverge with d, gradient step equations (3) imply:

$$q_{a/w}^{(1)} = q_\sigma + \tilde{q}_{a/w},$$

$$q_{a/w}^{(k+1)} = \max(q_{a/w}^{(k)}, q_{a/w}^{(1)} + \max(0, q_{w/a}^{(k)})). \quad (5)$$

We decompose our f as:

$$f_d^{(k)}(\mathbf{x}) = f_{d,\emptyset}^{(k)}(\mathbf{x}) + f_{d,a}^{(k)}(\mathbf{x}) + f_{d,w}^{(k)}(\mathbf{x}) + f_{d,aw}^{(k)}(\mathbf{x}), \quad (6)$$

$$f_{d,\emptyset}^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^{d}\hat{a}_r^{(0)}\phi'(\ldots)\hat{\mathbf{w}}_r^{(0),T}\mathbf{x},$$

$$f_{d,a}^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^{d}\delta\hat{a}_r^{(k)}\phi'(\ldots)\hat{\mathbf{w}}_r^{(0),T}\mathbf{x},$$

$$f_{d,w}^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^{d}\hat{a}_r^{(0)}\phi'(\ldots)\delta\hat{\mathbf{w}}_r^{(k),T}\mathbf{x},$$

$$f_{d,aw}^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^{d}\delta\hat{a}_r^{(k)}\phi'(\ldots)\delta\hat{\mathbf{w}}_r^{(k),T}\mathbf{x},$$

where $\phi'(\ldots)$ is a shorthand for $\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x})$ here. We assume $f_d^{(k)}(\mathbf{x}) \propto d^{q_f^{(k)}}$, $f_{d,\emptyset}^{(k)}(\mathbf{x}) \propto d^{q_{f,\emptyset}^{(k)}}$, and so on.

By definition of decomposition (6) terms, we have:

$$q_{f,\emptyset}^{(k)} = q_\sigma + \varkappa_\emptyset^{(k)}, \quad q_{f,a/w}^{(k)} = q_{a/w}^{(k)} + q_\sigma + \varkappa_{a/w}^{(k)},$$

$$q_{f,aw}^{(k)} = q_a^{(k)} + q_w^{(k)} + q_\sigma + \varkappa_{aw}^{(k)}, \quad (7)$$

where all $\varkappa \in [1/2, 1]$.

Our goal now is to compute $\varkappa$-terms for different values of q_σ and $\tilde{q}_{a/w}$. However it is more convenient to consider different cases for $q_a^{(1)}$ and $q_w^{(1)}$ instead.

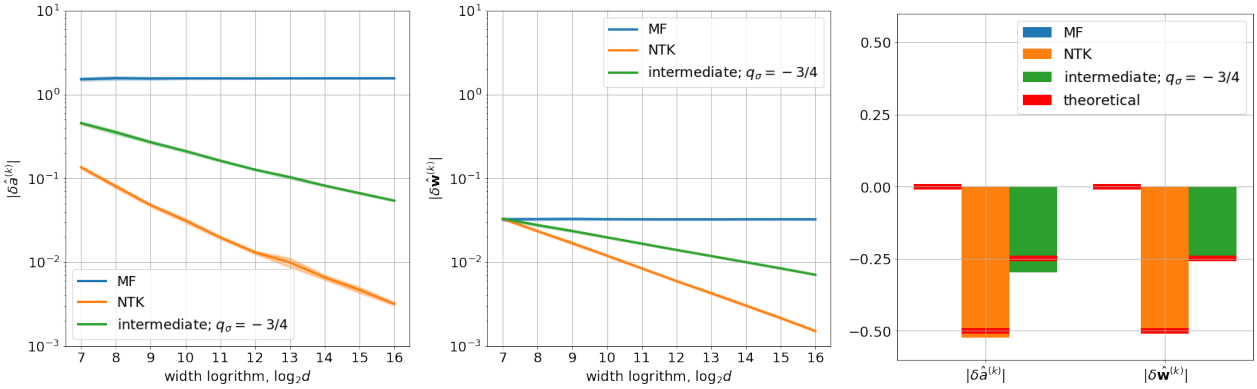

Figure 1. **Weight increments obey power-law dependencies with respect to the width.** *Left:* absolute output weight increments averaged over hidden neurons as functions of width d. *Center:* same for input weight increments. As one can see, weight increments are very well fitted with power-laws. *Right:* numerical estimates for exponents of corresponding power-laws, as well as their theoretical values (denoted by red ticks). As one can see, theoretical values match numerical estimates very well. *Setup:* We train a 1-hidden layer net on a subset of CIFAR2 (a dataset of first two classes of CIFAR10) of size 1000 with gradient descent. We take a reference net of width $d^* = 2^7 = 128$ trained with unscaled reference learning rates $\eta_a^* = \eta_w^* = 0.02$ and scale its hyperparameters according to MF (blue curves), NTK (orange curves) and intermediate scalings with $q_\sigma = -3/4$ (green curves, see main text). See SM A for further details.

D.1. $q_a^{(1)}$ and $q_w^{(1)}$ are both negative

In this case equations (5) imply $q_{a/w}^{(k)} = q_{a/w}^{(1)} < 0$ $\forall k \geq 1$. Hence $\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x}) \sim \phi'(\hat{\mathbf{w}}_r^{(0),T}\mathbf{x})$ as $d \to \infty$. Hence by the Central Limit Theorem, $\sum_{r=1}^d \hat{a}_r^{(0)}\phi'(\ldots)\hat{\mathbf{w}}_r^{(0),T}\mathbf{x} \propto d^{1/2}$. This means that $\varkappa_\emptyset^{(k)} = 1/2$.

At the same time, using the definition of the gradient step for $\delta\hat{\mathbf{w}}_r^{(k)}$,

$$f_{d,w}^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^d \hat{a}_r^{(0)}\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x})\delta\hat{\mathbf{w}}_r^{(k),T}\mathbf{x} \propto$$

$$\propto \hat{\eta}_w\sigma^2\sum_{r=1}^d \hat{a}_r^{(0)}\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x})(\hat{a}_r^{(0)} + \delta\hat{a}_r^{(k-1)}) \times$$

$$\times\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k-1)})^T\mathbf{x})\mathbf{x}^T\mathbf{x} \sim$$

$$\sim \hat{\eta}_w\sigma^2\sum_{r=1}^d (\hat{a}_r^{(0)})^2(\phi'(\hat{\mathbf{w}}_r^{(0),T}\mathbf{x}))^2\mathbf{x}^T\mathbf{x}.$$

We see that expression inside the sum has non-zero expectation, hence the sum scales as d, not as $d^{1/2}$. Hence $\varkappa_w^{(k)} = 1$. Using the similar reasoning we conclude that $\varkappa_a^{(k)} = 1$. For $f_{d,aw}^{(k)}$ we have:

$$f_{d,aw}^{(k)}(\mathbf{x}) = \sigma\sum_{r=1}^d \delta\hat{a}_r^{(k)}\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x})\delta\hat{\mathbf{w}}_r^{(k),T}\mathbf{x} \propto$$

$$\propto \hat{\eta}_a\hat{\eta}_w\sigma^3\sum_{r=1}^d (\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k-1)})^T\mathbf{x} \times$$

$$\times(\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k-1)})^T\mathbf{x}))^2 \times$$

$$\times\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x})(\hat{a}_r^{(0)} + \delta\hat{a}_r^{(k-1)})\mathbf{x}^T\mathbf{x} \sim$$

$$\sim \hat{\eta}_a\hat{\eta}_w\sigma^3\sum_{r=1}^d \hat{a}_r^{(0)}\mathbf{x}^T\mathbf{x}(\phi'(\hat{\mathbf{w}}_r^{(0),T}\mathbf{x}))^3\hat{\mathbf{w}}_r^{(0),T}\mathbf{x}.$$

Here all random terms of the sum has zero expectation and $\hat{a}_r^{(0)}$ is independent from $(\phi'(\hat{\mathbf{w}}_r^{(0),T}\mathbf{x}))^3\hat{\mathbf{w}}_r^{(0)}$; hence the sum scales as $d^{1/2}$ and consequently $\varkappa_{aw}^{(k)} = 1/2$.

Summing up, if $q_{a/w}^{(1)} < 0$, then $\varkappa_\emptyset^{(k)} = \varkappa_{aw}^{(k)} = 1/2$ and $\varkappa_{a/w}^{(k)} = 1$ $\forall k \geq 1$. Note that the NTK scaling is a subcase of this case.

D.2. $q_a^{(1)}$ and $q_w^{(1)}$ are both zeros

In this case equations (5) imply $q_{a/w}^{(k)} = q_{a/w}^{(1)} = 0$ $\forall k \geq 1$. Hence, generally, both $\delta\hat{a}^{(k)}$ and $\delta\hat{\mathbf{w}}^{(k)}$ depend on both $\hat{a}^{(0)}$ and $\hat{\mathbf{w}}^{(0)}$. This implies that sums in definitions of $f_{d,a}^{(k)}$, $f_{d,w}^{(k)}$ and $f_{d,aw}^{(k)}$ scale as d; hence $\varkappa_a^{(k)} = \varkappa_w^{(k)} = \varkappa_{aw}^{(k)} = 1$ $\forall k > 1$. Moreover, this implies that the sum

$$f_{d,\emptyset}^{(k)} = \sigma\sum_{r=1}^d \hat{a}_r^{(0)}\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta\hat{\mathbf{w}}_r^{(k)})^T\mathbf{x})\hat{\mathbf{w}}_r^{(0),T}\mathbf{x}$$

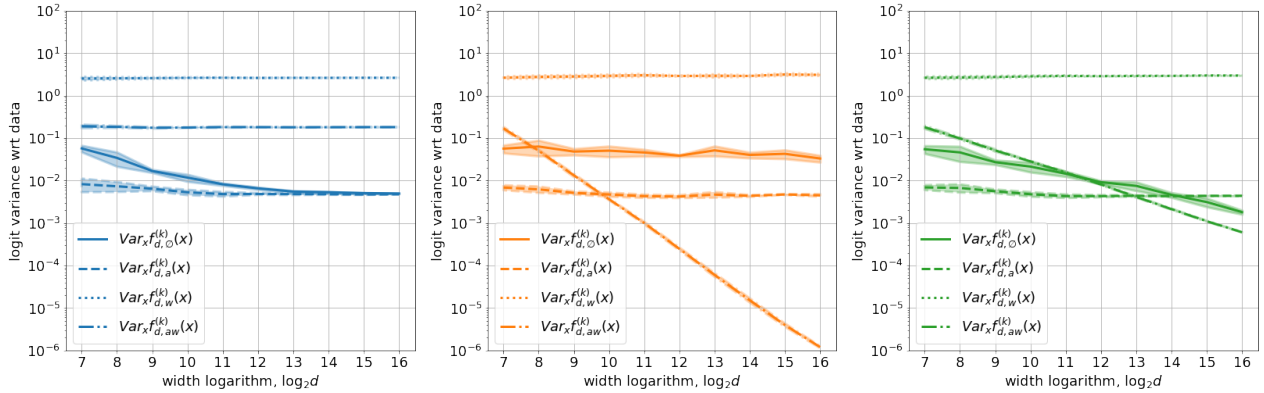

Figure 2. **Model decomposition terms obey power-law dependencies with respect to the width.** *Left:* the variance with respect to the data distribution for terms of model decomposition (6) as a function of width d for the mean-field scaling. *Center:* same for the NTK scaling. *Right:* same for the intermediate scaling with $q_\sigma = -3/4$. As one can see, the data distribution variance of model decomposition terms are well-fitted with power-laws. *Setup:* We train a 1-hidden layer net on a subset of CIFAR2 (a dataset of first two classes of CIFAR10) of size 1000 with a gradient descent. We take a reference net of width $d^* = 2^7 = 128$ trained with unscaled reference learning rates $\eta_a^* = \eta_w^* = 0.02$ and scale its hyperparameters according to MF (blue curves), NTK (orange curves) and intermediate scalings with $q_\sigma = -3/4$ (green curves, see text). See SM A for further details.

also scales as d. Hence $\varkappa_\emptyset^{(k)} = 1 \; \forall k \geq 1$. Note that this is the case of the MF scaling.

E. Other meaningful scalings

In the main text we have considered two solution classes for a system of equations and inequlaities that defines a meaningful scaling. One class corresponds to the case of both $q_a^{(1)}$ and $q_w^{(1)}$ being less than zero, while the other one corresponds to the case of both of them being zeros. In this section we consider all other possible cases.

E.1. $q_a^{(1)} = 0$, while $q_w^{(1)} < 0$

In this case equations (5) imply $q_a^{(k)} = q_a^{(1)} = 0$ and $q_w^{(k)} = q_w^{(1)} < 0 \; \forall k \geq 1$. Since $\hat{\mathbf{w}}^{(k)} \to \hat{\mathbf{w}}^{(0)}$, $\delta\hat{a}^{(k)}$ does not become independent on $\hat{\mathbf{w}}^{(0)}$ as $d \to \infty$; hence $\varkappa_a^{(k)} = 1$. Also, since $q_w^{(k)} < 0$, $\phi'(\hat{\mathbf{w}}^{(k),T}\mathbf{x}) \to \phi'(\hat{\mathbf{w}}^{(0),T}\mathbf{x})$; hence $\varkappa_\emptyset^{(k)} = 1/2$.

A condition $q_{f,a}^{(k)} = q_\sigma + q_a^{(1)} + \varkappa_a^{(k)} \leq 0$ then implies that $q_\sigma \leq -1$. Hence $q_{f,\emptyset}^{(k)} = q_\sigma + \varkappa_\emptyset^{(k)} \leq -1/2 < 0$. Moreover, $q_{f,w}^{(k)} = q_\sigma + q_w^{(k)} + \varkappa_w^{(k)} < 0$, since $\varkappa_w^{(k)} \leq 1$, and similarly, $q_{f,aw}^{(k)} = q_\sigma + q_a^{(k)} + q_w^{(k)} + \varkappa_{aw}^{(k)} < 0$ since $\varkappa_{aw}^{(k)} \leq 1$.

Hence in order to have a non-vanishing limit model we have to have $q_{f,a}^{(k)} = 0$ which implies $q_\sigma = -1$. Note that $q_a^{(1)} = q_\sigma + \tilde{q}_a = 0$; since then $\tilde{q}_a = 1$. Since $f_{d,a}^{(k)}$ is the only term of the model decomposition that remains finite as $d \to \infty$, we essentialy learn the output layer only in the limit of $d \to \infty$. Hence we can describe the dynamics of the limit model both in terms of the evolution of the limit measure and in terms of a constant deterministic limit kernel.

Indeed, suppose $\sigma = \sigma^* d^{-1}$ and $\hat{\eta}_a = \hat{\eta}_a^* d$. The limit measure evolution writes as follows:

$$f_\infty^{(k)}(\mathbf{x}) = \sigma^* \int \hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x})\, \mu_\infty^{(k)}(d\hat{a}, d\hat{\mathbf{w}});$$

$$\mu_\infty^{(k+1)} = \mathcal{T}_a(\mu_\infty^{(k)};\; \hat{\eta}_a^*\sigma^*, \sigma^*), \qquad \mu_\infty^{(0)} = \mathcal{N}_{1+d_0}(0, I),$$

where a gradient descent step operator $\mathcal{T}_a$ is defined on probabilistic measures μ supported on a finite set of atoms d as follows:

$$\mathcal{T}_a(\mu_d;\; \hat{\eta}_a^*\sigma^*, \sigma^*) = \frac{1}{d}\sum_{r=1}^{d} \delta_{\hat{a}'_r} \otimes \delta_{\hat{\mathbf{w}}_r},$$

where

$$\hat{a}'_r = \hat{a}_r - \hat{\eta}_a^*\sigma^* \mathbb{E}_{\mathbf{x},y} \left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f_d(\mathbf{x};\sigma^*)} \phi(\hat{\mathbf{w}}_r^T\mathbf{x}),$$

and $f_d(\mathbf{x};\sigma^*) = \sigma^* \int \hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x})\, \mu_d(d\hat{a}, d\hat{\mathbf{w}})$ for $(\hat{a}_r, \hat{\mathbf{w}}_r)$, $r \in [d]$, being atoms of measure μ_d.

Consider now a kernel $\tilde{\Theta}_{a,\infty}$ defined as follows:

$$\tilde{\Theta}_{a,\infty}(\mathbf{x}, \mathbf{x}') = \hat{\eta}_a^{*}\sigma^{*,2} \mathbb{E}_{\hat{\mathbf{w}}\sim\mathcal{N}(0,I_{d_0})} \phi(\hat{\mathbf{w}}^T\mathbf{x})\phi(\hat{\mathbf{w}}^T\mathbf{x}').$$

Using the same argument as in SM B, we can write a continuous-time evolution of the limit model in terms of this kernel:

$$\dot{f}_\infty^{(t)}(\mathbf{x}') = -\mathbb{E}_{\mathbf{x},y} \left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f_\infty^{(t)}(\mathbf{x})} \tilde{\Theta}_{a,\infty}(\mathbf{x}, \mathbf{x}'),$$

$$f_\infty^{(0)}(\mathbf{x}) = \mathbb{E}_{(\hat{a},\hat{\mathbf{w}})\sim\mathcal{N}(0,I_{1+d_0})}\hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x}) = 0.$$

Moreover, for the same argument as in SM B, the similar evolution equation holds also for the discrete-time evolution:

$$\Delta f_\infty^{(k)}(\mathbf{x}') = -\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f_\infty^{(k)}(\mathbf{x})}\tilde{\Theta}_{a,\infty}(\mathbf{x},\mathbf{x}').$$

E.2. $q_w^{(1)} = 0$, while $q_a^{(1)} < 0$

This case is almost analogous to the previous one. Equations (5) imply $q_w^{(k)} = q_w^{(1)} = 0$ and $q_a^{(k)} = q_a^{(1)} < 0\ \forall k \geq 1$, and $\delta\hat{w}^{(k)}$ does not become independent on $\hat{a}^{(0)}$ as $d \to \infty$; hence $\varkappa_w^{(k)} = 1$. Note that in contrast to the previous case, since $q_w^{(k)} = 0$, $\phi'(\hat{\mathbf{w}}^{(k),T}\mathbf{x}) \nrightarrow \phi'(\hat{\mathbf{w}}^{(0),T}\mathbf{x})$; hence $\varkappa_\emptyset^{(k)} = 1$.

A condition $q_{f,w}^{(k)} = q_\sigma + q_w^{(1)} + \varkappa_w^{(k)} \leq 0$ (or, equivalently, a condition $q_{f,\emptyset}^{(k)} = q_\sigma + \varkappa_\emptyset^{(k)} \leq 0$) then implies that $q_\sigma \leq -1$. Hence $q_{f,a}^{(k)} = q_\sigma + q_a^{(k)} + \varkappa_a^{(k)} < 0$, since $\varkappa_a^{(k)} \leq 1$, and similarly, $q_{f,aw}^{(k)} = q_\sigma + q_a^{(k)} + q_w^{(k)} + \varkappa_{aw}^{(k)} < 0$, since $\varkappa_{aw}^{(k)} \leq 1$.

Hence in order to have a non-vanishing limit model we have to have $q_{f,\emptyset}^{(k)} = q_{f,w}^{(k)} = 0$, which implies $q_\sigma = -1$. Note that $q_w^{(1)} = q_\sigma + \tilde{q}_w = 0$; since then $\tilde{q}_w = 1$. In this case we again essentialy learn only a single layer in the limit of $d \to \infty$. However a kernel which drives the dynamics evolves with k since $w^{(k)} \nrightarrow w^{(0)}$:

$$\tilde{\Theta}_{w,\infty}^{(k)}(\mathbf{x},\mathbf{x}') = \hat{\eta}_w^*\sigma^{*,2}\lim_{d\to\infty}\frac{1}{d}\sum_{d=1}^{\infty}\mathbb{E}_{\hat{a}\sim\mathcal{N}(0,1)}|\hat{a}|^2\times \\ \times\phi'(\hat{\mathbf{w}}^{(k),T}\mathbf{x})\phi'(\hat{\mathbf{w}}^{(k),T}\mathbf{x}')\mathbf{x}^T\mathbf{x}'.$$

Nevertheless, the dynamics can be described in terms of the measure evolution similar to the previous case.

E.3. $q_a^{(1)} > 0$, while $q_a^{(1)} + q_w^{(1)} \leq 0$

In this case equations (5) imply $q_a^{(k)} = q_a^{(1)} > 0$, while $q_w^{(k)} = q_a^{(1)} + q_w^{(1)} \leq 0\ \forall k > 1$. Similar to the case of SM E.1, $\delta\hat{a}^{(k)}$ does not become independent on $\hat{\mathbf{w}}^{(0)}$ as $d \to \infty$; hence $\varkappa_a^{(k)} = 1$.

Consider $k > 1$. A condition $q_{f,a}^{(k)} = q_\sigma + q_a^{(1)} + \varkappa_a^{(k)} \leq 0$ then implies $q_\sigma \leq -1 - q_a^{(1)} < -1$. Hence $q_{f,\emptyset}^{(k)} = q_\sigma + \varkappa_\emptyset^{(k)} < 0$ since $\varkappa_\emptyset^{(k)} \leq 1$. Moreover, $q_{f,w}^{(k)} = q_\sigma + q_w^{(k)} + \varkappa_w^{(k)} < 0$ since $\varkappa_w^{(k)} \leq 1$ and $q_w^{(k)} = q_a^{(1)} + q_w^{(1)} \leq 0$, and similarly, $q_{f,aw}^{(k)} = q_\sigma + q_a^{(k)} + q_w^{(k)} + \varkappa_{aw}^{(k)} \leq q_{f,a}^{(k)} \leq 0$ since $\varkappa_{aw}^{(k)} \leq 1$.

Hence in order to have a non-vanishing limit model we have to have $q_{f,a}^{(k)} = 0$, which implies $q_a^{(1)} = q_\sigma + \tilde{q}_a = -1 - q_\sigma$. Since then $\tilde{q}_a = -1 - 2q_\sigma$, while $q_\sigma < -1$. Suppose $q_w^{(k)} = q_a^{(1)} + q_w^{(1)} < 0$. In this case $q_{f,aw}^{(k)} < 0$, hence $f_{d,a}^{(k)}$ is the only term of the model decomposition that remains finite as $d \to \infty$, and we learn the output layer only in the limit of $d \to \infty$, as was the case of SM E.1. In this case we are again able to describe the dynamics of the limit model both in terms of the evolution of the limit measure and in terms of a constant deterministic limiting kernel.

While the kernel description does not change at all compared to the case described in SM E.1, measure evolution equations require slight modifications. Indeed, suppose $\sigma = \sigma^* d^{q_\sigma}$ and $\hat{\eta}_a = \hat{\eta}_a^* d^{-1-2q_\sigma}$. The limit measure evolution writes as follows:

$$f_\infty^{(k)}(\mathbf{x}) = \sigma^*\int \hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x})\,\mu_\infty^{(k)}(d\hat{a},d\hat{\mathbf{w}});$$

$$\mu_\infty^{(k+1)} = \mathcal{T}_a(\mu_\infty^{(k)};\ \hat{\eta}_a^*\sigma^*,\sigma^*), \qquad \mu_\infty^{(0)} = \delta\otimes\mathcal{N}(0,I_{d_0}),$$

where a gradient descent step operator $\mathcal{T}_a$ is defined on probabilistic measures μ supported on a finite set of atoms d as follows:

$$\mathcal{T}_a(\mu_d;\ \hat{\eta}_a^*\sigma^*,\sigma^*) = \frac{1}{d}\sum_{r=1}^{d}\delta_{\hat{a}_r'}\otimes\delta_{\hat{\mathbf{w}}_r},$$

where

$$\hat{a}_r' = \hat{a}_r - \hat{\eta}_a^*\sigma^*\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial\ell(y,z)}{\partial z}\right|_{z=f_d(\mathbf{x};\sigma^*)}\phi(\hat{\mathbf{w}}_r^T\mathbf{x}),$$

and $f_d(\mathbf{x};\sigma^*) = \sigma^*\int\hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x})\,\mu_d(d\hat{a},d\hat{\mathbf{w}})$ for $(\hat{a}_r,\hat{\mathbf{w}}_r)$, $r\in[d]$, being atoms of measure μ_d.

The only thing changed here is that in the limit output weights $\hat{a}$ are initialized with zeros. The reason for this is that the increment at the first step $\delta\hat{a}^{(0)} = -\hat{\eta}_a\sigma\mathbb{E}\,\nabla_f^{(0)}\ell\,\phi(\hat{\mathbf{w}}^{(0),T}\mathbf{x})$ grows as d^{-1-q_σ} as $d\to\infty$. Hence $\hat{a}^{(k)} \to \delta\hat{a}^{(k)}$ as $d\to\infty$ for $k\geq 1$.

Suppose now $q_w^{(k)} = q_a^{(1)} + q_w^{(1)} = 0$. In this case $\delta\hat{a}^{(k)}$ and $\delta\hat{\mathbf{w}}^{(k)}$ do not become independent, since $\hat{\mathbf{w}}^{(k)} \nrightarrow \hat{\mathbf{w}}^{(0)}$; hence $\varkappa_{aw}^{(k)} = 1$. This implies that $q_{f,aw}^{(k)} = q_\sigma + q_a^{(k)} + q_w^{(k)} + \varkappa_{aw}^{(k)} = 0$ for $k > 1$, hence two terms of the model decomposition remain finite as $d\to\infty$: $f_{d,a}^{(k)}$ and $f_{d,aw}^{(k)}$. Note that $q_a^{(1)} + q_w^{(1)} = 0$ implies $\tilde{q}_w = -\tilde{q}_a - 2q_\sigma = 1$.

Suppose $\hat{\eta}_w = \hat{\eta}_w^* d$. In this case we are again able to describe the dynamics of the limit model in terms of the evolution of the limit measure:

$$f_\infty^{(k)}(\mathbf{x}) = \sigma^*\int \hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x})\,\mu_\infty^{(k)}(d\hat{a},d\hat{\mathbf{w}});$$

$$\mu_\infty^{(k+1)} = \mathcal{T}_a(\mu_\infty^{(k)};\ \hat{\eta}_a^*\sigma^*,\sigma^*), \qquad \mu_\infty^{(0)} = \delta\otimes\mathcal{N}(0,I_{d_0}),$$

where a gradient descent step operator $\mathcal{T}_a$ is defined on probabilistic measures μ supported on a finite set of atoms d as follows:

$$\mathcal{T}(\mu_d;\ \hat{\eta}_a^*\sigma^*, \hat{\eta}_w^*\sigma^*, \sigma^*) = \frac{1}{d}\sum_{r=1}^{d} \delta_{\hat{a}_r'} \otimes \delta_{\hat{\mathbf{w}}_r},$$

where

$$\hat{a}_r' = \hat{a}_r - \hat{\eta}_a^*\sigma^*\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f_d(\mathbf{x};\sigma^*)} \phi(\hat{\mathbf{w}}_r^T\mathbf{x}),$$

$$\hat{\mathbf{w}}_r' = \hat{\mathbf{w}}_r - \hat{\eta}_w^*\sigma^*\mathbb{E}_{\mathbf{x},y}\left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f_d(\mathbf{x};\sigma^*)} \hat{a}_r\phi'(\hat{\mathbf{w}}_r^T\mathbf{x}),$$

and $f_d(\mathbf{x};\sigma^*) = \sigma^*\int \hat{a}\phi(\hat{\mathbf{w}}^T\mathbf{x})\,\mu_d(d\hat{a}, d\hat{\mathbf{w}})$ for $(\hat{a}_r, \hat{\mathbf{w}}_r)$, $r \in [d]$, being atoms of measure μ_d.

We have a zero initialization for the output weights for the same reason as for the case of $q_w^{(k)} < 0$. Note that in contrast to the above-mentioned case, the case of $q_w^{(k)} = 0$ cannot be described in terms of a constant limit kernel. Indeed, we have a stochastic time-dependent kernel for finite width d associated with output weights learning:

$$\tilde{\Theta}_{a,\infty}^{(k)}(\mathbf{x},\mathbf{x}') = \hat{\eta}_a^*\sigma^{*,2}\frac{1}{d}\sum_{r=1}^{d}\phi(\hat{\mathbf{w}}_r^{(k),T}\mathbf{x})\phi(\hat{\mathbf{w}}_r^{(k),T}\mathbf{x}').$$

This kernel converges to a deterministic one as $d \to \infty$ by the Law of Large Numbers, however, the limit kernel stays step-dependent, since $\hat{\mathbf{w}}^{(k)} = \hat{\mathbf{w}}^{(0)} + \delta\hat{\mathbf{w}}^{(k)}$, while the last term here does not vanish as $d \to \infty$.

Note that the "default" case we have considered in the main text falls into the current case. Indeed, by default we have $\sigma \propto d^{-1/2}$ and $\eta_{a/w} \propto 1$. This implies $q_\sigma = -1/2$, $\tilde{q}_a = 1$ and $\tilde{q}_w = 0$; consequently, $q_a^{(1)} = 1/2$ and $q_w^{(1)} = -1/2$. However, as we have shown above, having $q_\sigma \leq -1 - q_a^{(1)} = -3/2$ is necessary to guarantee that the limit model does not diverge. As we observe in Figure 1 of the main text a limit model resulted from the default scaling indeed diverges.

E.4. $q_w^{(1)} > 0$, while $q_a^{(1)} + q_w^{(1)} \leq 0$

The difference between this case and the previous one is essentially the same as between cases of SM E.2 and of SM E.1. For this reason we leave this case as an exercise for the reader.

E.5. $q_a^{(1)} + q_w^{(1)} > 0$

Suppose first that $q_a^{(1)} > 0$. In this case equations (5) imply $q_w^{(2)} = q_a^{(1)} + q_w^{(1)} > 0$ and $q_a^{(2)} \geq q_a^{(1)} > 0$. It is easy to see that equations 5 further imply $q_a^{(2k)} = q_w^{(2k)} = k(q_a^{(2)} + q_w^{(2)})$ $\forall k \geq 1$. This means that $q_a^{(k)}$ and $q_w^{(k)}$ grow linearly with k. Hence all of $q_{f,a}^{(k)}$, $q_{f,w}^{(k)}$, $q_{f,aw}^{(k)}$ become positive for large enough k irrespective of q_σ.

Obviously, the same holds if $q_w^{(1)} > 0$. Hence in this case our analysis suggests that a limit model $f_\infty^{(k)}$ diverges with d for large enough k. However, when our analysis predicts that a limit model diverges, we cannot guarantee that $\nabla_f^{(k)}\ell$ does not vanish with d, hence equations 5 become generally incorrect. Indeed, if a model reaches 100% train accuracy at step k, then $\nabla_f^{(k)}\ell$ vanishes exponentially if f grows. This means that f cannot really diverge width d if it reaches 100% train accuracy.

F. A discrete-time mean-field limit of a network with a single hidden layer

In this section we omit "hats" for brevity, assuming all relevant quantities to be scaled appropriately.

Recall that in the MF limit $\sigma \propto d^{-1}$ and $\eta_{a/w} \propto d$. Suppose $\sigma = \sigma^* d^{-1}$ and w.l.o.g. $\eta_{a/w} = \eta^* d$.

We closely follow the measure-theoretic formalism of Sirignano & Spiliopoulos (2020). Consider a measure in $(a, \mathbf{w})$-space at each step k for a given d:

$$\mu_d^{(k)} = \frac{1}{d}\sum_{r=1}^{d}\delta_{a_r^{(k)}} \otimes \delta_{\mathbf{w}_r^{(k)}}.$$

Given this, a neural network output can be represented as follows:

$$f_d^{(k)}(\mathbf{x}) = \sigma^*\int a\phi(\mathbf{w}^T\mathbf{x})\mu_d^{(k)}(da,\, d\mathbf{w}).$$

A gradient descent step is written as follows:

$$\Delta a_r^{(k)} = -\eta^*\sigma^*\mathbb{E}_{\mathbf{x},y}\nabla_f^{(k)}\ell\ \phi(\mathbf{w}_r^{(k),T}\mathbf{x}),$$

$$\Delta\mathbf{w}_r^{(k)} = -\eta^*\sigma^*\mathbb{E}_{\mathbf{x},y}\nabla_f^{(k)}\ell\ a_r^{(k)}\phi'(\mathbf{w}_r^{(k),T}\mathbf{x})\mathbf{x}. \quad (8)$$

For technical reasons we assume weights a_r and $\mathbf{w}_r$ $\forall r \in [d]$ to be initialized from the distribution $\mathcal{P}$ with compact support:

$$a_r^{(0)} \sim \mathcal{P}, \quad w_{r,j}^{(0)} \sim \mathcal{P} \quad \forall r \in [d]\ \forall j \in [d_0]. \quad (9)$$

One can notice that in the main body of this work we have assumed $\mathcal{P}$ to be $\mathcal{N}(0,1)$ that does not have a compact support. Nevertheless, it is more common in practice to use a truncated normal distribution instead of the original normal one, which was used in the main body for the ease of explanation only.

We introduce a transition operator $\mathcal{T}$ which represents a gradient descent step (8):

$$\mu_d^{(k+1)} = \mathcal{T}(\mu_d^{(k)};\ \eta^*, \sigma^*). \tag{10}$$

This operator depends explicitly on σ^* because $\nabla_f^{(k)}\ell$ is a gradient of $f_d^{(k)}$ and the latter depends on σ^*. This representation clearly shows that a gradient descent defines a Markov chain for measures on the weight space with deterministic transitions. The initial measure $\mu_d^{(0)}$ is given by initial conditions (9). Since they are random, measure $\mu_d^{(k)}$ is a random measure for any $k \geq 0$ and for any $d \in \mathbb{N}$. Nevertheless, for all $k \geq 0$ $\mu_d^{(k)}$ converges to a corresponding limit measure as the following theorem states:

Theorem 1. *Suppose $\ell(y,\cdot) \in C^2(\mathbb{R})$, $\partial\ell(y,z)/\partial z$ is bounded and Lipschitz continuous and ϕ is Lipschitz continuous. Suppose also that $\mathbf{x}$ has finite moments up to the fourth one. Finally, assume that the distribution of initial weights $\mathcal{P}$ has compact support. Then $\forall k \geq 0$ there exists a measure $\mu_\infty^{(k)}$ such that $\mu_d^{(k)}$ converges to $\mu_\infty^{(k)}$ weakly as $d \to \infty$ wrt to the 2-Wasserstein metric and each measure $\mu_d^{(k)}$ is supported on a ball $\mathcal{B}_{R_k}$ a.s. for all d.*

Proof. We prove this by induction on k.

Let $k = 0$. Any measure μ on the weight space is uniquely determined by its action on all $g \in C(\mathbb{R}^{1+d_0})$ with compact support: $\langle g, \mu\rangle = \int g(a, \mathbf{w})\mu(da,\, d\mathbf{w})$. If this measure is random, then the last integral is a random variable. Hence $\mu_d^{(0)}$ converges to $\mu_\infty^{(0)} = \mathcal{P}$ weakly as $d \to \infty$, iff for all $g \in C(\mathbb{R}^{1+d_0})$ with compact support $\langle g, \mu_d^{(0)}\rangle$ converges to $\langle g, \mu_\infty^{(0)}\rangle$ weakly as $d \to \infty$.

Let $h \in C_b(\mathbb{R})$. Consider

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h\left(\left\langle g, \mu_d^{(0)}\right\rangle\right) =$$
$$= \lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h\left(\frac{1}{d}\sum_{r=1}^{d} g(a_r^{(0)}, \mathbf{w}_r^{(0)})\right) =$$
$$= h\left(\mathbb{E}_{a^{(0)},\mathbf{w}^{(0)}} g\left(a^{(0)}, \mathbf{w}^{(0)}\right)\right) = h\left(\left\langle g, \mu_\infty^{(0)}\right\rangle\right),$$

where the second equality comes from the Law of Large Numbers which is valid since initial weights are i.i.d. This proves a weak convergence of $\langle g, \mu_d^{(0)}\rangle$ to $\langle g, \mu_\infty^{(0)}\rangle$. As was noted above, this is equivalent to a weak convergence of measures $\mu_d^{(0)}$:

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mu_d^{(0)}] = h[\mu_\infty^{(0)}]$$

for any $h \in C_b(\mathcal{M}(\mathbb{R}^{1+d_0}))$.

Also, since all $a_r \sim \mathcal{P}$, $w_{r,j} \sim \mathcal{P}$ and $\mathcal{P}$ has compact support, $\mu_d^{(0)}$ has compact support almost surely. Hence we can write $\mu_d^{(0)} \in \mathcal{M}(\mathcal{B}_{R_0}^{1+d_0})$ a.s. for some $R_0 < \infty$ $\forall d$.

We have proven the induction base. By induction assumption, we have $\mu_d^{(k)} \in \mathcal{M}(\mathcal{B}_{R_k}^{1+d_0})$ a.s. for some $R_k < \infty$ $\forall d$. Let for any $h \in C_b(\mathcal{M}(\mathbb{R}^{1+d_0}))$

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mu_d^{(k)}] = h[\mu_\infty^{(k)}].$$

By definition, this means weak convergence of measures $\mu_d^{(k)}$ to $\mu_\infty^{(k)}$. Then we have:

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mu_d^{(k+1)}] = \lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mathcal{T}(\mu_d^{(k)})].$$

In order to prove that this limit exists and equals to $h[\mathcal{T}(\mu_\infty^{(k)})]$ we have to show that $h \circ \mathcal{T} \in C_b(\mathcal{M}(\mathcal{B}_{R_k}^{1+d_0}))$.

We prove the following lemma in Section F.1:

Lemma 1. *Given conditions of Theorem 1 and $R < \infty$, the transition operator T that performs a gradient descent step (10) is continuous wrt the 2-Wasserstein metric on $\mathcal{M}(\mathcal{B}_R^{1+d_0})$.*

Hence $h \circ \mathcal{T} \in C_b(\mathcal{M}(\mathcal{B}_{R_k}^{1+d_0}))$. Since then, by the induction hypothesis for all $h \in C_b(\mathcal{M}(\mathbb{R}^{1+d_0}))$

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mu_d^{(k+1)}] =$$
$$= \lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mathcal{T}(\mu_d^{(k)})] = h[\mathcal{T}(\mu_\infty^{(k)})].$$

We then define $\mu_\infty^{(k+1)} = \mathcal{T}(\mu_\infty^{(k)})$.

Also, it easy to see that since ϕ, ϕ' and $\partial\ell(y,z)/\partial z$ are bounded and the distribution of $\mathbf{x}$ has a bounded variation, $\mu_d^{(k)} \in \mathcal{M}(\mathcal{B}_{R_k}^{1+d_0})$ a.s. implies $\mu_d^{(k+1)} = \mathcal{T}\mu_d^{(k)} \in \mathcal{M}(\mathcal{B}_{R_{k+1}}^{1+d_0})$ a.s. for some $R_{k+1} < \infty$.

We have proven that for all $k \geq 0$ $\mu_d^{(k)}$ converges to $\mu_\infty^{(k)}$ weakly as $d \to \infty$ wrt the 2-Wasserstein metric and $\mu_d^{(k)}$ has compact support a.s. for any $d \in \mathbb{N}$. □

Corollary 1 (Theorem 1 of Section 3, restated)**.** *Given the same conditions as in Theorem 1, following statements hold:*

1. *$\forall k \geq 0$ $\mu_d^{(k)}$ converges to $\mu_\infty^{(k)}$ in probability as $d \to \infty$;*
2. *$f_d^{(k)}(\mathbf{x})$ converges to some $f_\infty^{(k)}(\mathbf{x})$ in probability as $d \to \infty$ $\forall \mathbf{x} \in \mathcal{X}$.*

Proof. Since weak convergence to a constant implies convergence in probability, the first statement directly follows from Theorem 1.

By definition, weak convergence of $\mu_d^{(k)}$ means for any $h \in C_b(\mathcal{M}(\mathbb{R}^{1+d_0}))$

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} h[\mu_d^{(k)}] = h[\mu_\infty^{(k)}].$$

Hence for any $g \in C_b(\mathbb{R})$

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} g(f_d^{(k)}(\mathbf{x})) = \\ = \lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} g(f[\mu_d^{(k)};\mathbf{x}]) = \\ = \lim_{d\to\infty} \mathbb{E}_{\mathbf{a}^{(0)},W^{(0)}} (g\circ f)[\mu_d^{(k)};\mathbf{x}] = (g\circ f)[\mu_\infty^{(k)};\mathbf{x}],$$

since $f[\cdot;\mathbf{x}] \in C(\mathcal{M}(\mathbb{R}^{1+d_0}))$ for any $\mathbf{x} \in \mathcal{X}$.

Hence $f_d^{(k)}(\mathbf{x}) = f[\mu_d^{(k)};\mathbf{x}]$ converges weakly to $f_\infty^{(k)}(\mathbf{x}) = f[\mu_\infty^{(k)};\mathbf{x}]$ as $d \to \infty$. By the same argument as above, this implies convergence in probability. □

F.1. A gradient descent step defines a continuous operator in the space of weight-space measures

Proof of Lemma 1. Without loss of generality assume $\sigma^* = \eta^* = 1$. Consider a sequence of measures $\mu_d \in \mathcal{M}(\mathcal{B}_R^{1+d_0})$ that converges to $\mu_\infty \in \mathcal{M}(\mathcal{B}_R^{1+d_0})$ wrt the 2-Wasserstein metric. We have to prove that $\mathcal{T}\mu_d$ converges to $\mathcal{T}\mu_\infty$ wrt the 2-Wasserstein metric.

Define $\theta_d = (a_d, \mathbf{w}_d) \in \mathcal{B}_R^{1+d_0}$ and $\delta\theta_d = \theta_\infty - \theta_d = (a_\infty - a_d, \mathbf{w}_\infty - \mathbf{w}_d) \in \mathcal{B}_R^{1+d_0}$. For a given d consider a sequence of measures $\mu_{d,\infty}^j \in \mathcal{M}(\mathcal{B}_R^{1+d_0} \otimes \mathcal{B}_R^{1+d_0})$ with marginals equal to μ_d and μ_∞ respectively, as required by the definition of the Wasserstein metric. Choose a sequence in such a way that

$$\lim_{j\to\infty} \int (\|\delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) = \\ = \inf_{\mu_{d,\infty}} \int (\|\delta\theta_d\|_2^2 \, \mu_{d,\infty}(d\theta_d,\, d\theta_\infty) = \mathcal{W}_2^2(\mu_d, \mu_\infty),$$

where infium is taken over all $\mu_{d,\infty} \in \mathcal{M}(\mathcal{B}_R^{1+d_0} \otimes \mathcal{B}_R^{1+d_0})$ with marginals equal to μ_d and μ_∞ respectively as required by the definition of the Wasserstein metric. A sequence $\{\mu_{d,\infty}^j\}_{j=1}^\infty$ exists by properties of infium. Then we have the following:

$$\mathcal{W}_2^2(\mathcal{T}\mu_d, \mathcal{T}\mu_\infty) \leq \\ \leq \lim_{j\to\infty} \int \|\delta\theta_d + \delta\Delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty),$$

where we have defined

$$\Delta\theta_d = \Big(-\mathbb{E}\, \nabla_{f_d}\ell \, \phi(\mathbf{w}_d^T\mathbf{x}), -\mathbb{E}\, \nabla_{f_d}\ell \, a_d \phi'(\mathbf{w}_d^T\mathbf{x})\mathbf{x}\Big),$$

$$\nabla_{f_d}\ell = \left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f[\mu_d;\mathbf{x}]}$$

and $\delta\Delta\theta_d = \Delta\theta_\infty - \Delta\theta_d$ respectively. From this follows:

$$\mathcal{W}_2^2(\mathcal{T}\mu_d, \mathcal{T}\mu_\infty) \leq \lim_{j\to\infty} \int \|\delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) + \\ + \lim_{j\to\infty} \int \|\delta\Delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) + \\ +2 \lim_{j\to\infty} \int \langle \delta\theta_d, \delta\Delta\theta_d\rangle \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty).$$

Consequently,

$$\mathcal{W}_2^2(\mathcal{T}\mu_d, \mathcal{T}\mu_\infty) \leq \mathcal{W}_2^2(\mu_d, \mu_\infty) + \tag{11}$$

$$+ \lim_{j\to\infty} \int \|\delta\Delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) + \tag{12}$$

$$+4R \lim_{j\to\infty} \sqrt{\int \|\delta\Delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty)}. \tag{13}$$

The last term comes (1) from the Cauchy-Schwartz inequality: $\langle \delta\theta_d, \delta\Delta\theta_d\rangle \leq \|\delta\theta_d\|_2 \|\delta\Delta\theta_d\|_2$, (2) from the fact that both μ_d and μ_∞ are concentrated in a ball of radius R: $\|\delta\theta_d\|_2 = \|\theta_d - \theta_\infty\|_2 \leq \|\theta_d\|_2 + \|\theta_\infty\|_2 \leq 2R$, and (3) from Jensen's inequality: $\int \|\theta\|_2 \, \mu(d\theta) \leq \sqrt{\int \|\theta\|_2^2 \, \mu(d\theta)}$, for μ being a probability measure.

The first term converges to zero by the definition of the sequence of measures μ_d. Consider the second term:

$$\int \|\delta\Delta\theta_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) = \\ = \int (\delta\Delta a_d)^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) + \tag{14} \\ + \int \|\delta\Delta\mathbf{w}_d\|_2^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty).$$

Consider then the first term here:

$$\int (\delta\Delta a_d)^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) = \\ = \int \Big(\mathbb{E}_{\mathbf{x},y}\Big(\nabla_{f_d}\ell \, \phi(\mathbf{w}_d^T\mathbf{x}) - \\ -\nabla_{f_\infty}\ell \, \phi(\mathbf{w}_\infty^T\mathbf{x})\Big)\Big)^2 \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) = \\ = \int (\mathbb{E}_{\mathbf{x},y}(g(\mathbf{x},\theta_d)h(\mathbf{x},y,\mu_d) - \\ -g(\mathbf{x},\theta_\infty)h(\mathbf{x},y,\mu_\infty)))^2 \, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty),$$

where we have defined

$$g(\mathbf{x},\theta) = g(\mathbf{x},(a,\mathbf{w})) = \phi(\mathbf{w}^T\mathbf{x}),$$

$$h(\mathbf{x},y,\mu) = \left.\frac{\partial \ell(y,z)}{\partial z}\right|_{z=f[\mu;\mathbf{x}]}.$$

W.l.o.g. assume ϕ has a Lipschitz constant 1: $\phi(\cdot) \in \mathrm{Lip}(\mathbb{R};1)$. From this follows that $g(\mathbf{x},\cdot) \in$

$\mathrm{Lip}(\mathbb{R}^{1+d_0}; \|\mathbf{x}\|_2)$. It is easy to see that since we consider measures supported on $\mathcal{B}_R$, $f[\cdot, \mathbf{x}] \in \mathrm{Lip}(\mathcal{M}(\mathcal{B}_R^{1+d_0}); 2R\|\mathbf{x}\|_2)$ wrt the 2-Wasserstein metric. Indeed,

$$\begin{aligned}
|f[\mu_d, \mathbf{x}] - f[\mu_\infty, \mathbf{x}]| = \\
= \Big| \int a_d \phi(\mathbf{w}_d^T \mathbf{x})\, \mu(da_d,\, d\mathbf{w}_d) - \\
- \int a_\infty \phi(\mathbf{w}_\infty^T \mathbf{x})\, \mu(da_\infty,\, d\mathbf{w}_\infty) \Big| = \\
= \Big| \int (a_d \phi(\mathbf{w}_d^T \mathbf{x}) - a_\infty \phi(\mathbf{w}_\infty^T \mathbf{x}))\, \mu(d\theta_d)\, \mu(d\theta_\infty) \Big| \leq \\
\leq \int |a_d \phi(\mathbf{w}_d^T \mathbf{x}) - a_\infty \phi(\mathbf{w}_\infty^T \mathbf{x})|\, \mu(d\theta_d)\, \mu(d\theta_\infty) \leq \\
\leq \int (|a_d||\phi(\mathbf{w}_d^T \mathbf{x}) - \phi(\mathbf{w}_\infty^T \mathbf{x})| + \\
+ |a_d - a_\infty||\phi(\mathbf{w}_\infty^T \mathbf{x})|)\, \mu(d\theta_d)\, \mu(d\theta_\infty) \leq \\
\leq R\|\mathbf{x}\|_2 \int (\|\delta \mathbf{w}_d\|_2 + |\delta a_d|)\, \mu(d\theta_d)\, \mu(d\theta_\infty) \leq \\
\leq R\|\mathbf{x}\|_2 \sqrt{\int \|\mathbf{w}_d - \mathbf{w}_\infty\|_2^2\, \mu(d\theta_d)\, \mu(d\theta_\infty)} + \\
+ R\|\mathbf{x}\|_2 \sqrt{\int |a_d - a_\infty|^2\, \mu(d\theta_d)\, \mu(d\theta_\infty)} \leq \\
\leq 2R\|\mathbf{x}\|_2 \mathcal{W}_2(\mu_d, \mu_\infty),
\end{aligned}$$

where we have used Jensen's inequality: $\int \|\theta\|_2\, \mu(d\theta) \leq \sqrt{\int \|\theta\|_2^2\, \mu(d\theta)}$ since μ is a probability measure.

W.l.o.g. $\partial \ell / \partial z \in \mathrm{Lip}(\mathbb{R}; 1)\ \forall y \in \{0, 1\}$. Hence the latter implies $h(\mathbf{x}, y, \cdot) \in \mathrm{Lip}(\mathcal{M}(\mathcal{B}_R^{1+d_0}); 2R\|\mathbf{x}\|_2)$.

Taking into account that w.l.o.g. $\partial \ell / \partial z$ and ϕ' are bounded by 1, we have:

$$\begin{aligned}
|g(\mathbf{x}, \theta_d) h(\mathbf{x}, y, \mu_d) - g(\mathbf{x}, \theta_\infty) h(\mathbf{x}, y, \mu_\infty)| \leq \\
\leq |g(\mathbf{x}, \theta_d) - g(\mathbf{x}, \theta_\infty)| + \\
+ R\|\mathbf{x}\|_2 |h(\mathbf{x}, y, \mu_d) - h(\mathbf{x}, y, \mu_\infty)| \leq \\
\leq \|\mathbf{x}\|_2 \|\theta_d - \theta_\infty\|_2 + 2R^2 \|\mathbf{x}\|_2^2 \mathcal{W}_2(\mu_d, \mu_\infty).
\end{aligned}$$

From this follows:

$$\begin{aligned}
(\mathbb{E}_{\mathbf{x},y}(g(\mathbf{x}, \theta_d) h(\mathbf{x}, y, \mu_d) - g(\mathbf{x}, \theta_\infty) h(\mathbf{x}, y, \mu_\infty)))^2 \leq \\
\leq \mathbb{E}_{\mathbf{x},y}(g(\mathbf{x}, \theta_d) h(\mathbf{x}, y, \mu_d) - g(\mathbf{x}, \theta_\infty) h(\mathbf{x}, y, \mu_\infty))^2 \leq \\
\leq \mathbb{E}_{\mathbf{x},y} \|\mathbf{x}\|_2^2 \|\theta_d - \theta_\infty\|_2^2 + 4R^4 \mathbb{E}_{\mathbf{x},y} \|\mathbf{x}\|_2^4 \mathcal{W}_2^2(\mu_d, \mu_\infty) + \\
+ 4R^2 \mathbb{E}_{\mathbf{x},y} \|\mathbf{x}\|_2^3 \|\theta_d - \theta_\infty\|_2 \mathcal{W}_2(\mu_d, \mu_\infty).
\end{aligned}$$

Hence

$$\begin{aligned}
\lim_{j \to \infty} \int (\mathbb{E}_{\mathbf{x},y}(g(\mathbf{x}, \theta_d) h(\mathbf{x}, y, \mu_d) - \\
- g(\mathbf{x}, \theta_\infty) h(\mathbf{x}, y, \mu_\infty)))^2\, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) \leq \\
\leq \mathbb{E}_{\mathbf{x},y} \|\mathbf{x}\|_2^2 \mathcal{W}_2^2(\mu_d, \mu_\infty) + 4R^4 \mathbb{E}_{\mathbf{x},y} \|\mathbf{x}\|_2^4 \mathcal{W}_2^2(\mu_d, \mu_\infty) + \\
+ 4R^2 \mathbb{E}_{\mathbf{x},y} \|\mathbf{x}\|_2^3 \mathcal{W}_2^2(\mu_d, \mu_\infty) = \\
= \mathbb{E}_{\mathbf{x},y}(\|\mathbf{x}\|_2 + 2R^2 \|\mathbf{x}\|_2^2)^2 \mathcal{W}_2^2(\mu_d, \mu_\infty).
\end{aligned}$$

We can apply the same logic to the second term of (14) to get the same upper bound:

$$\begin{aligned}
\int (\delta \Delta \mathbf{w}_d)^2\, \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) = \\
= \int \Big(\mathbb{E}_{\mathbf{x},y} \left(\nabla_{f_d} \ell\, a_d \phi'(\mathbf{w}_d^T \mathbf{x}) - \right. \\
\left. - \nabla_{f_\infty} \ell\, a_\infty \phi'(\mathbf{w}_\infty^T \mathbf{x}) \right) \Big)^2 \mu_{d,\infty}^j(d\theta_d,\, d\theta_\infty) \leq \\
\leq \mathbb{E}_{\mathbf{x},y}(\|\mathbf{x}\|_2 + 2R^2 \|\mathbf{x}\|_2^2)^2 \mathcal{W}_2^2(\mu_d, \mu_\infty).
\end{aligned}$$

Applying this upper bound to equation (13), we finally get the following:

$$\begin{aligned}
\lim_{d \to \infty} \mathcal{W}_2^2(\mathcal{T}\mu_d, \mathcal{T}\mu_\infty) \leq \lim_{d \to \infty} \Big(\mathcal{W}_2^2(\mu_d, \mu_\infty) + \\
+ 2\mathbb{E}_{\mathbf{x},y}(\|\mathbf{x}\|_2 + 2R^2 \|\mathbf{x}\|_2^2)^2 \mathcal{W}_2^2(\mu_d, \mu_\infty) + \\
+ 4R \sqrt{2\mathbb{E}_{\mathbf{x},y}(\|\mathbf{x}\|_2 + 2R^2 \|\mathbf{x}\|_2^2)^2}\, \mathcal{W}_2(\mu_d, \mu_\infty) \Big) = 0,
\end{aligned}$$

where the last equality is valid, because by assumptions $\mathbf{x}$ has finite moments up to the fourth one. Hence $\mathcal{T}\mu_d$ converges to $\mathcal{T}\mu_\infty$ wrt the 2-Wasserstein metric.

Summing up, we have proven that $\mathcal{T}$ is continuous wrt the 2-Wasserstein metric.

□

G. The mean-field limit is trivial for the case of more than two hidden layers

Here we re-write the definition of a multi-layer net, as well as the gradient descent step on scaled quantities:

$$f(\mathbf{x}; \mathbf{a}, V^{1:H}, W) = \sum_{r_H=1}^{d} a_{r_H} \phi(f_{r_H}^H(\mathbf{x}; V^{1:H}, W)),$$

where

$$f_{r_{h+1}}^{h+1}(\mathbf{x}; V^{1:h+1}, W) = \sum_{r_h=1}^{d} v_{r_{h+1} r_h}^{h+1} \phi(f_{r_h}^h(\mathbf{x}; V^{1:h}, W)),$$

$$f_{r_0}^0(\mathbf{x}, W) = \mathbf{w}_{r_0}^T \mathbf{x}.$$

The gradient descent step:

$$\Delta\hat{a}_{r_H}^{(k)} = -\hat{\eta}_a \sigma^{H+1}\mathbb{E}\,\nabla_f^{(k)}\ell\;\phi(\hat{f}_{r_H}^{H,(k)}(\mathbf{x})),$$

$$\Delta\hat{v}_{r_H r_{H-1}}^{H,(k)} = -\hat{\eta}_{v^H} \sigma^{H+1}\mathbb{E}\,\nabla_f^{(k)}\ell\;\hat{a}_{r_H}^{(k)}\phi(\hat{f}_{r_{H-1}}^{H-1,(k)}(\mathbf{x})),$$

$$\dots$$

$$\Delta\hat{\mathbf{w}}_{r_0}^{(k)} = -\hat{\eta}_w \sigma^{H+1}\mathbb{E}\,\nabla_f^{(k)}\ell\sum_{r_H=1}^{d}\hat{a}_{r_H}^{(k)}\phi'(\hat{f}_{r_H}^{H,(k)}(\mathbf{x}))\times$$

$$\times\sum_{r_{H-1}=1}^{d}\hat{v}_{r_H r_{H-1}}^{H,(k)}\phi'(\hat{f}_{r_{H-1}}^{H-1,(k)}(\mathbf{x}))\times\dots$$

$$\dots\times\sum_{r_1=1}^{d}\hat{v}_{r_2 r_1}^{2,(k)}\phi'(\hat{f}_{r_1}^{1,(k)}(\mathbf{x}))\hat{v}_{r_1 r_0}^{1,(k)}\phi'(\hat{\mathbf{w}}_{r_0}^{(k),T}\mathbf{x})\mathbf{x}.$$

$$\hat{a}_{r_H}^{(0)}\sim\mathcal{N}(0,I),\;\hat{v}_{r_h r_{h-1}}^{h,(0)}\sim\mathcal{N}(0,I),\;\hat{\mathbf{w}}_{r_0}^{(0)}\sim\mathcal{N}(0,I), \tag{15}$$

where we have denoted $\hat{f}_{r_h}^{h,(k)}(\mathbf{x}) = f_{r_h}^h(\mathbf{x};\hat{V}^{(k),1:h},\hat{W}^{(k)})$.

Similarly to the case of $H=0$ (see Section 3), we consider a power-law dependence on d for σ and learning rates, as a result introducing q_σ, $\tilde{q}_a$, $\tilde{q}_{v^h}$ and $\tilde{q}_w$. In Section 4 we have shown that for the mean-field limit we should have $q_\sigma=-1$, $\tilde{q}_{a/w}=1$ and $\tilde{q}_{v^h}=2$.

We now show that for $H\geq 2$ the mean-field limit is trivial: $\lim_{d\to\infty} f_d^{(k)}(\mathbf{x}) = 0$. Similarly to the case of $H=0$, we introduce weight increments $\delta\hat{a}_{r_H}^{(k)} = \hat{a}_{r_H}^{(k)} - \hat{a}_{r_H}^{(0)}$, $\delta\hat{v}_{r_h r_{h-1}}^{h,(k)} = \hat{v}_{r_h r_{h-1}}^{h,(k)} - \hat{v}_{r_h r_{h-1}}^{h,(0)}$ and $\delta\hat{\mathbf{w}}_{r_0}^{(k)} = \hat{\mathbf{w}}_{r_0}^{(k)} - \hat{\mathbf{w}}_{r_0}^{(0)}$, and assume a power-law dependence on d for them resulting in the introduction of exponents $q_a^{(k)}$, $q_{v^h}^{(k)}$ and $q_w^{(k)}$.

Analogically to a single hidden layer case, we decompose our f:

$$f_d^{(k)}(\mathbf{x}) = f_{d,\emptyset}^{(k)}(\mathbf{x}) + f_{d,a}^{(k)}(\mathbf{x}) + \sum_{h=1}^{H} f_{d,v^h}^{(k)}(\mathbf{x}) + f_{d,w}^{(k)}(\mathbf{x}) + \\ + \dots + f_{d,av^{1:H}w}^{(k)}(\mathbf{x}), \tag{16}$$

where the exact definition of each term can be derived from its sub-index: e.g. $f_{d,wa}^{(k)}$ has $\delta\hat{a}^{(k)}$, $\delta\hat{\mathbf{w}}^{(k)}$ and $\hat{v}^{h,(0)}\;\forall h\in[H]$ terms.

Introducing an exponent q for each term, we get:

$$q_f^{(k)} = \max(q_{f,\emptyset}^{(k)}, q_{f,a}^{(k)}, \dots, q_{f,av^{1:H}w}^{(k)}), \quad q_f^{(0)} = 2q_\sigma + 1. \tag{17}$$

We write all of the terms of the decomposition for f in a unified way. Let Θ_h be a subset of $\{a, v^{1:H}, w\}$ of size h. Then:

$$q_{f,\Theta_h}^{(k)} = H(\varkappa_{\Theta_h}^{(k)} + q_\sigma) + \sum_{\theta\in\Theta_h} q_\theta^{(k)}, \tag{18}$$

where $\varkappa_{\Theta_h}^{(k)}\in[1/2,1]$ comes from the same logic as in the single hidden layer case. Since $q_\sigma=-1$, if we show that all $q_\theta^{(k)} < 0\;\forall k\geq 1$, then we conclude that all components of decomposition (16) vanish.

Let us look on the gradient descent dynamics (15). It implies the following equalities for $k=0$:

$$q_{a/w}^{(1)} = \tilde{q}_{a/w} + (H+1)q_\sigma + \frac{H}{2} = -\frac{H}{2}, \tag{19}$$

$$q_{v^h}^{(1)} = \tilde{q}_{v^h} + (H+1)q_\sigma + \frac{H-1}{2} = -\frac{H-1}{2},$$

which come from the fact that all $\hat{a}^{(0)}$, $\hat{v}^{h,(0)}$ and $\hat{\mathbf{w}}^{(0)}$ are independent and $\propto 1$. Indeed, gradient updates for $\delta\hat{a}$ and $\delta\hat{\mathbf{w}}$ have H sums each, and each sum scales as $d^{1/2}$ (this where the term $H/2$ comes from); at the same time gradient updates for $\delta\hat{v}^h$ have $H-1$ sums each.

Due to the symmetry of the gradient step dynamics, $q_{v^1}^{(1)} = \dots = q_{v^H}^{(1)}$ imply $q_{v^1}^{(k)} = \dots = q_{v^H}^{(k)}\;\forall k\geq 1$. We shall denote it with $q_v^{(k)}$ then.

Suppose $H\geq 2$. We prove that $q_{a/w}^{(k)} \leq q_{a/w}^{(1)} = -H/2$ and $q_v^{(k)} \leq q_v^{(1)} = (1-H)/2\;\forall k\geq 1$ by induction. The induction base $k=1$ is trivial. For the sake of illustration, we first consider the induction step for q_w:

$$q_w^{(k+1)} \leq \max\Bigg(q_w^{(k)}, \tilde{q}_w + (H+1)q_\sigma + \\ + \max\Bigg(\frac{H}{2}, \frac{H+1}{2} + q_a^{(k)}, \frac{H+1}{2} + q_v^{(k)}, \\ H + q_a^{(k)} + q_v^{(k)}, H + 2q_v^{(k)}\Bigg)\Bigg) \leq \\ \leq \max\Bigg(-\frac{H}{2}, -\frac{H}{2} + \max\Bigg(0, \frac{1}{2} + q_a^{(k)}, \frac{1}{2} + q_v^{(k)}, \\ \frac{H}{2} + q_a^{(k)} + q_v^{(k)}, \frac{H}{2} + 2q_v^{(k)}\Bigg)\Bigg) \leq \\ \leq \max\Bigg(-\frac{H}{2}, -\frac{H}{2} + \max\Bigg(0, \frac{1-H}{2}, \frac{2-H}{2}, \\ \frac{1-H}{2}, \frac{2-H}{2}\Bigg)\Bigg) = -\frac{H}{2}. \tag{20}$$

All inequalities except the first come from the induction hypothesis. We now demonstrate where the first inequality

comes from. Recall that $\|\delta\hat{\mathbf{w}}^{(k+1)}\| \propto d^{q_w^{(k+1)}}$ and

$$\delta\hat{\mathbf{w}}_{r_0}^{(k+1)} = \delta\hat{\mathbf{w}}_{r_0}^{(k)} -$$
$$- \hat{\eta}_w \sigma^{H+1} \mathbb{E}\, \nabla_f^{(k)} \ell \sum_{r_H=1}^{d} (\hat{a}_{r_H}^{(0)} + \delta\hat{a}_{r_H}^{(k)}) \phi'(\hat{f}_{r_H}^{H,(k)}(\mathbf{x})) \times$$
$$\times \sum_{r_{H-1}=1}^{d} (\hat{v}_{r_H r_{H-1}}^{H,(0)} + \delta\hat{v}_{r_H r_{H-1}}^{H,(k)}) \phi'(\hat{f}_{r_{H-1}}^{H-1,(k)}(\mathbf{x})) \times \dots$$
$$\dots \times \sum_{r_1=1}^{d} (\hat{v}_{r_2 r_1}^{2,(0)} + \delta\hat{v}_{r_2 r_1}^{2,(k)}) \phi'(\hat{f}_{r_1}^{1,(k)}(\mathbf{x})) \times$$
$$\times (\hat{v}_{r_1 r_0}^{1,(0)} + \delta\hat{v}_{r_1 r_0}^{1,(k)}) \phi'((\hat{\mathbf{w}}_{r_0}^{(0)} + \delta\hat{\mathbf{w}}_{r_0}^{(k)})^T \mathbf{x}) \mathbf{x}. \quad (21)$$

Here we have a product of H sums, by expanding which we obtain a sum of 2^{H+1} products of sums in total; for example, for $H = 2$ we have:

$$\sum_{r_2=1}^{d} (\hat{a}_{r_2}^{(0)} + \delta\hat{a}_{r_2}^{(k)}) \phi'(\hat{f}_{r_2}^{2,(k)}(\mathbf{x})) \times$$
$$\times \sum_{r_1=1}^{d} (\hat{v}_{r_2 r_1}^{2,(0)} + \delta\hat{v}_{r_2 r_1}^{2,(k)}) \phi'(\hat{f}_{r_1}^{1,(k)}(\mathbf{x})) \times$$
$$\times (\hat{v}_{r_1 r_0}^{1,(0)} + \delta\hat{v}_{r_1 r_0}^{1,(k)}) \phi'((\hat{\mathbf{w}}_{r_0}^{(0)} + \delta\hat{\mathbf{w}}_{r_0}^{(k)})^T \mathbf{x}) \mathbf{x} =$$
$$= \sum_{r_2=1}^{d} \hat{a}_{r_2}^{(0)} \phi'(\dots) \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\dots) \hat{v}_{r_1 r_0}^{1,(0)} \phi'(\dots) \mathbf{x} +$$
$$+ \sum_{r_2=1}^{d} \delta\hat{a}_{r_2}^{(k)} \phi'(\dots) \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\dots) \hat{v}_{r_1 r_0}^{1,(0)} \phi'(\dots) \mathbf{x} +$$
$$+ \sum_{r_2=1}^{d} \hat{a}_{r_2}^{(0)} \phi'(\dots) \sum_{r_1=1}^{d} \delta\hat{v}_{r_2 r_1}^{2,(k)} \phi'(\dots) \hat{v}_{r_1 r_0}^{1,(0)} \phi'(\dots) \mathbf{x} +$$
$$+ \sum_{r_2=1}^{d} \hat{a}_{r_2}^{(0)} \phi'(\dots) \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\dots) \delta\hat{v}_{r_1 r_0}^{1,(k)} \phi'(\dots) \mathbf{x} + \dots$$
$$\dots + \sum_{r_2=1}^{d} \delta\hat{a}_{r_2}^{(k)} \phi'(\dots) \sum_{r_1=1}^{d} \delta\hat{v}_{r_2 r_1}^{2,(k)} \phi'(\dots) \delta\hat{v}_{r_1 r_0}^{1,(k)} \phi'(\dots) \mathbf{x} =$$
$$= \Sigma_{d,\emptyset}^{(k)} + \Sigma_{d,a}^{(k)} + \Sigma_{d,v^1}^{(k)} + \Sigma_{d,v^2}^{(k)} +$$
$$+ \Sigma_{d,v^1 v^2}^{(k)} + \Sigma_{d,v^2 a}^{(k)} + \Sigma_{d,a v^1}^{(k)} + \Sigma_{d,a v^1 v^2}^{(k)},$$

where the notation we have introduced is intuitive: for example, $\Sigma_{d,av^1}^{(k)} = \sum_{r_2=1}^{d} \delta\hat{a}_{r_2}^{(0)} \phi'(\dots) \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\dots) \delta\hat{v}_{r_1 r_0}^{1,(k)} \phi'(\dots) \mathbf{x}$.

If we assume power-law dependencies for all Σ-terms, i.e. $\Sigma_{d,\emptyset}^{(k)} \propto d^{q_{\Sigma,\emptyset}^{(k)}}$, $\Sigma_{d,a}^{(k)} \propto d^{q_{\Sigma,a}^{(k)}}$ and so on, using heuristic rules mentioned in Section 3, from (21) we get the following:

$$q_w^{(k+1)} = \max(q_w^{(k)}, \tilde{q}_w + (H+1) q_\sigma +$$
$$+ \max(q_{\Sigma,\emptyset}^{(k)}, q_{\Sigma,a}^{(k)}, q_{\Sigma,v^1}^{(k)}, q_{\Sigma,v^2}^{(k)}, \dots, q_{\Sigma,av^1v^2}^{(k)})).$$

First consider $\Sigma_{d,av^1v^2}^{(k)}$. This term is a product of two sums with d terms each. Since each sum cannot grow faster than d, we get the following upper bound:

$$q_{\Sigma,av^1v^2}^{(k)} \le q_a^{(k)} + 2q_v^{(k)} + 2.$$

Similar upper bounds hold for all other Σ-terms; in particular, we have:

$$q_{\Sigma,v^1v^2}^{(k)} \le 2q_v^{(k)} + 2, \ \max(q_{\Sigma,v^2a}^{(k)}, q_{\Sigma,av^1}^{(k)}) \le q_a^{(k)} + q_v^{(k)} + 2.$$

For $\Sigma_{d,\emptyset}^{(k)}$ we compute the corresponding exponent exactly: $q_{\Sigma,\emptyset}^{(k)} = 1$. In this case both sums are the sums of asymptotically independent terms with zero mean. Indeed, we have:

$$\Sigma_{d,\emptyset}^{(k)} = \sum_{r_2=1}^{d} \hat{a}_{r_2}^{(0)} \phi'(\hat{f}_{r_2}^{2,(k)}(\mathbf{x})) \times$$
$$\times \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\hat{f}_{r_1}^{1,(k)}(\mathbf{x})) \hat{v}_{r_1 r_0}^{1,(0)} \phi'((\hat{\mathbf{w}}_{r_0}^{(0)} + \delta\hat{\mathbf{w}}_{r_0}^{(k)})^T \mathbf{x}) \mathbf{x} \sim$$
$$\sim \sum_{r_2=1}^{d} \hat{a}_{r_2}^{(0)} \phi'(\hat{f}_{r_2}^{2,(0)}(\mathbf{x})) \times$$
$$\times \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\hat{f}_{r_1}^{1,(0)}(\mathbf{x})) \hat{v}_{r_1 r_0}^{1,(0)} \phi'(\hat{\mathbf{w}}_{r_0}^{(0),T} \mathbf{x}) \mathbf{x},$$

where the asymptotic equivalence takes place, because by the induction hypothesis $q_{a/w}^{(k)} \le -H/2 < 0$ and $q_v^{(k)} \le (1-H)/2 < 0$.

Finally, let us consider "linear" terms, i.e. $\Sigma_{d,a}^{(k)}$, $\Sigma_{d,v^1}^{(k)}$, $\Sigma_{d,v^2}^{(k)}$. We consider $\Sigma_{d,a}^{(k)}$ for simplicity; two other terms can be analysed in a similar manner. Here we have a similar asymptotic relation as we had for $\Sigma_{d,\emptyset}^{(k)}$:

$$\Sigma_{d,a}^{(k)} = \sum_{r_2=1}^{d} \delta\hat{a}_{r_2}^{(k)} \phi'(\hat{f}_{r_2}^{2,(k)}(\mathbf{x})) \times$$
$$\times \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\hat{f}_{r_1}^{1,(k)}(\mathbf{x})) \hat{v}_{r_1 r_0}^{1,(0)} \phi'((\hat{\mathbf{w}}_{r_0}^{(0)} + \delta\hat{\mathbf{w}}_{r_0}^{(k)})^T \mathbf{x}) \mathbf{x} \sim$$
$$\sim \sum_{r_2=1}^{d} \delta\hat{a}_{r_2}^{(k)} \phi'(\hat{f}_{r_2}^{2,(0)}(\mathbf{x})) \times$$
$$\times \sum_{r_1=1}^{d} \hat{v}_{r_2 r_1}^{2,(0)} \phi'(\hat{f}_{r_1}^{1,(0)}(\mathbf{x})) \hat{v}_{r_1 r_0}^{1,(0)} \phi'(\hat{\mathbf{w}}_{r_0}^{(0),T} \mathbf{x}) \mathbf{x}.$$

Let us now recall the gradient step for $\delta\hat{a}^{(k)}$:

$$\delta\hat{a}_{r_2}^{(k)} = \delta\hat{a}_{r_2}^{(k-1)} - \hat{\eta}_a \sigma^3 \mathbb{E}\, \nabla_f^{(k-1)} \ell \times$$
$$\times \phi\left(\sum_{r_1=1}^{d} (\hat{v}_{r_2 r_1}^{2,(0)} + \delta\hat{v}_{r_2 r_1}^{2,(k-1)}) \phi(\hat{f}_{r_1}^{1,(k-1)}(\mathbf{x})) \right).$$

Since by the induction hypothesis $q_v^{(k-1)} \leq (1-H)/2 < 0$, $\delta\hat{a}_{r_2}^{(k)}$ depends on $\hat{v}_{r_2 r_1}^{2,(0)}$, even as $d \to \infty$. This means that the sum over r_2 in the definition of $\Sigma_{d,a}^{(k)}$ above grows as d, while the sum over r_1 still grows as $d^{1/2}$, as was the case for $\Sigma_{d,\emptyset}^{(k)}$. Hence

$$q_{\Sigma,a/v^1/v^2}^{(k)} = q_{a/v^1/v^2}^{(k)} + 3/2.$$

Finally, for $H = 2$ we get the following:

$$\begin{aligned}
q_w^{(k+1)} &= \max(q_w^{(k)}, \tilde{q}_w + (H+1)q_\sigma + \\
&\quad + \max(q_{\Sigma,\emptyset}^{(k)}, q_{\Sigma,a}^{(k)}, q_{\Sigma,v^1}^{(k)}, q_{\Sigma,v^2}^{(k)}, \ldots, q_{\Sigma,av^1v^2}^{(k)})) \leq \\
&\leq \max(q_w^{(k)}, \tilde{q}_w + (H+1)q_\sigma + \max(1, q_a^{(k)} + 3/2, q_v^{(k)} + 3/2, \\
&\quad q_a^{(k)} + q_v^{(k)} + 2, 2q_v^{(k)} + 2, q_a^{(k)} + 2q_v^{(k)} + 2)) = \\
&= \max(q_w^{(k)}, \tilde{q}_w + (H+1)q_\sigma + \\
&+ \max(1, q_a^{(k)} + 3/2, q_v^{(k)} + 3/2, q_a^{(k)} + q_v^{(k)} + 2, 2q_v^{(k)} + 2)),
\end{aligned}$$

where the last equality comes from the fact that $q_{a/v/w}^{(k)} < 0$ by the induction hypothesis. Directly extending this technique to the case of $H \geq 2$ results in the first inequality of (20).

Applying the similar technique to q_a and q_v we get the following:

$$\begin{aligned}
q_a^{(k+1)} &\leq \max\Bigg(q_a^{(k)}, \tilde{q}_a + (H+1)q_\sigma + \\
&\quad + \max\Bigg(\frac{H}{2}, \frac{H+1}{2} + q_w^{(k)}, \frac{H+1}{2} + q_v^{(k)}, \\
&\quad H + q_w^{(k)} + q_v^{(k)}, H + 2q_v^{(k)}\Bigg)\Bigg) \leq \\
&\leq \max\Bigg(-\frac{H}{2}, -\frac{H}{2} + \max\Bigg(0, \frac{1}{2} + q_w^{(k)}, \frac{1}{2} + q_v^{(k)}, \\
&\quad \frac{H}{2} + q_w^{(k)} + q_v^{(k)}, \frac{H}{2} + 2q_v^{(k)}\Bigg)\Bigg) \leq \\
&\leq \max\Bigg(-\frac{H}{2}, -\frac{H}{2} + \max\Bigg(0, \frac{1-H}{2}, \frac{2-H}{2}, \\
&\quad \frac{1-H}{2}, \frac{2-H}{2}\Bigg)\Bigg) = -\frac{H}{2},
\end{aligned}$$

$$\begin{aligned}
q_v^{(k+1)} &\leq \max\Bigg(q_v^{(k)}, \tilde{q}_v + (H+1)q_\sigma + \\
&\quad + \max\Bigg(\frac{H-1}{2}, \frac{H}{2} + q_a^{(k)}, \frac{H}{2} + q_w^{(k)}, \frac{H}{2} + q_v^{(k)}, \\
&H-1+q_a^{(k)}+q_w^{(k)}, H-1+q_w^{(k)}+q_v^{(k)}, H-1+q_a^{(k)}+q_v^{(k)}\Bigg)\Bigg) \leq \\
&\leq \max\Bigg(\frac{1-H}{2}, \frac{1-H}{2} + \max\Bigg(0, \frac{1}{2} + q_a^{(k)}, \frac{1}{2} + q_w^{(k)}, \\
&\quad \frac{1}{2} + q_v^{(k)}, \frac{H-1}{2} + q_a^{(k)} + q_w^{(k)}, \frac{H-1}{2} + q_w^{(k)} + q_v^{(k)}, \\
&\quad \frac{H-1}{2} + q_a^{(k)} + q_v^{(k)}\Bigg)\Bigg) \leq \\
&\leq \max\Bigg(\frac{1-H}{2}, \frac{1-H}{2} + \max\Bigg(0, \frac{1-H}{2}, \frac{1-H}{2}, \\
&\quad \frac{2-H}{2}, \frac{-1-H}{2}, -\frac{H}{2}, -\frac{H}{2}\Bigg)\Bigg) = \frac{1-H}{2},
\end{aligned}$$

for all $h \in [H]$. The difference between $q_{a/w}$ and q_v comes from the fact that the gradient step for $\delta\hat{v}^h$ has $H-1$ sums instead of H.

Summing up, we have proven by induction that $\forall k \geq 1$ $q_{a/w}^{(k)} \leq q_{a/w}^{(1)} = -H/2 < 0$ and $q_v^{(k)} \leq q_v^{(1)} = (1-H)/2 < 0$. Hence due to (18), $q_{f,\Theta_h}^{(k)} < 0$, hence all components of decomposition (16) vanish and $\lim_{d\to\infty} f_d^{(k)} = 0$.

H. Comparing scalings for small learning rates

As we have noted in Section 3, the MF limit provides the most accurate approximation for a finite-width reference network. However as we demonstrate here the NTK limit becomes the most accurate approximation for a finite-width reference network if learning rates are sufficiently small and the number of training steps is fixed.

Figure 3 shows results for two different setups: training a one hidden layer net with gradient descent for 50 epochs with reference learning rates $\eta_a^* = \eta_w^* = 0.02$ (the same setup as in Figure 1 of Section 3 of the main text) and the same setup but with $\eta_a^* = \eta_w^* = 0.0002$. As one can see, MF and intermediate limits do not preserve the variance of the CE loss but the NTK limit does.

In Section 3 we have argued that the MF limit provides a better approximation for a finite-width reference net, because all terms of decomposition (6) are preserved, however, as we have previously observed in SM C, the term $f_{d,\emptyset}^{(k)}$ is not strictly preserved but approaches a non-zero constant

for large d. As we observe in the right plot of the bottom row, the width $2^{16} = 65536$ is not yet enough for $f^{(k)}_{d,\emptyset}$ to reach its asymptotics for the MF limit if learning rates are small: see blue solid curve. Nevertheless, for large learning rates (right plot of the top row) this term does reach its asymptotics.

However one of the decomposition term vanishes for the NTK limit but for the MF limit it does not: $f^{(k)}_{d,aw}$. Let us rewrite the definition of this term here:

$$f^{(k)}_{d,aw}(\mathbf{x}) = \sigma \sum_{r=1}^{d} \delta \hat{a}^{(k)}_r \phi'(\dots) \delta \hat{\mathbf{w}}^{(k),T}_r \mathbf{x}.$$

This term depends quadratically on weight increments and each weight increment is proportional to a corresponding learning rate. Hence this term grows quadratically with learning rates. By the same logic, terms $f^{(k)}_{d,a}$ and $f^{(k)}_{d,w}$ grow linearly with learning rates and $f^{(k)}_{d,\emptyset}$ has no polynomial dependence on learning rates. This reasoning implies that the term $f^{(k)}_{d,aw}$ vanishes faster than others as learning rates go to zero. Hence the effect of non-preserving this term becomes negligible if learning rates are small. Because of this, the advantage of the MF limit over the NTK limit disappears for sufficiently small learning rates. This effect is clearly shown in the right column of Figure 3. For large learning rates (top row) the term $f^{(k)}_{d,aw}$ is the second-largest term in decomposition (6) of the reference network: see a dash-dot curve, however it becomes negligible for one hundred times smaller learning rates (bottom row).

I. Experiments for other setups

As was noted in Section A, in the present work we typically train a network using a full-batch gradient descent (or RMSProp) for 50 epochs (or equivalently, training steps) on a subset of CIFAR2 of size 1000. The reason for this is that our theory is developed for binary classification, it assumes exact gradient computations, and because training up to convergence is not necessary for our framework.

In this section we experiment with modifications of our usual setup: see Figure 4 for the case of a one hidden layer net trained with the (stochastic) gradient descent. The top row represents the usual case of the full-batch gradient descent training for 50 epochs with unscaled reference learning rates $\eta^*_a = \eta^*_w = 0.02$ applied to a subset of CIFAR2. For the next row we set the batch size to 100, while keeping the number of gradient updates. As we observe, applying a stochastic gradient descent instead of the full-batch one does not introduce any qualitative changes. For the third row we take a full CIFAR2 (with training size being 10000 instead of 1000), while keeping the batch size to be 1000. It is hard to spot any qualitative changes in this setup as well. For the bottom row we increase the number of epochs (training steps) by the factor of 10, while keeping the rest of the options. In this case all plots change, which is expected since 50 epochs of the original setup is not enough for convergence of training procedure. As we observe in the center column, in this case some of the terms of decomposition (6) do not obey power-laws but converge to power-laws for large d.

We also consider a multi-class classification instead of a binary one: see Figure 5. The top row corresponds to the usual scenario of a binary classification on a subset of CIFAR2 of size 1000; it is given for the reference. The middle row corresponds to the same scenario but on a subset of MNIST of size 1000; MNIST has 10 classes instead of two. Comparing these two scenarios does not reveal any qualitative changes.

The bottom row corresponds to the most realistic scenario among the ones we have considered. Here we train a one hidden layer network on a full MNIST dataset for 6000 gradient steps using a mini-batch gradient descent with batches of size 100. With this number of epochs the optimization process nearly converges. As we see, for this scenario the maximum width $d = 2^{16} = 65536$ we were able to afford was not enough to reach the asymptotic regime fully (center column). This is the reason for discrepancies between numerical estimates of exponents of decomposition (6) terms and their theoretical values (right column).

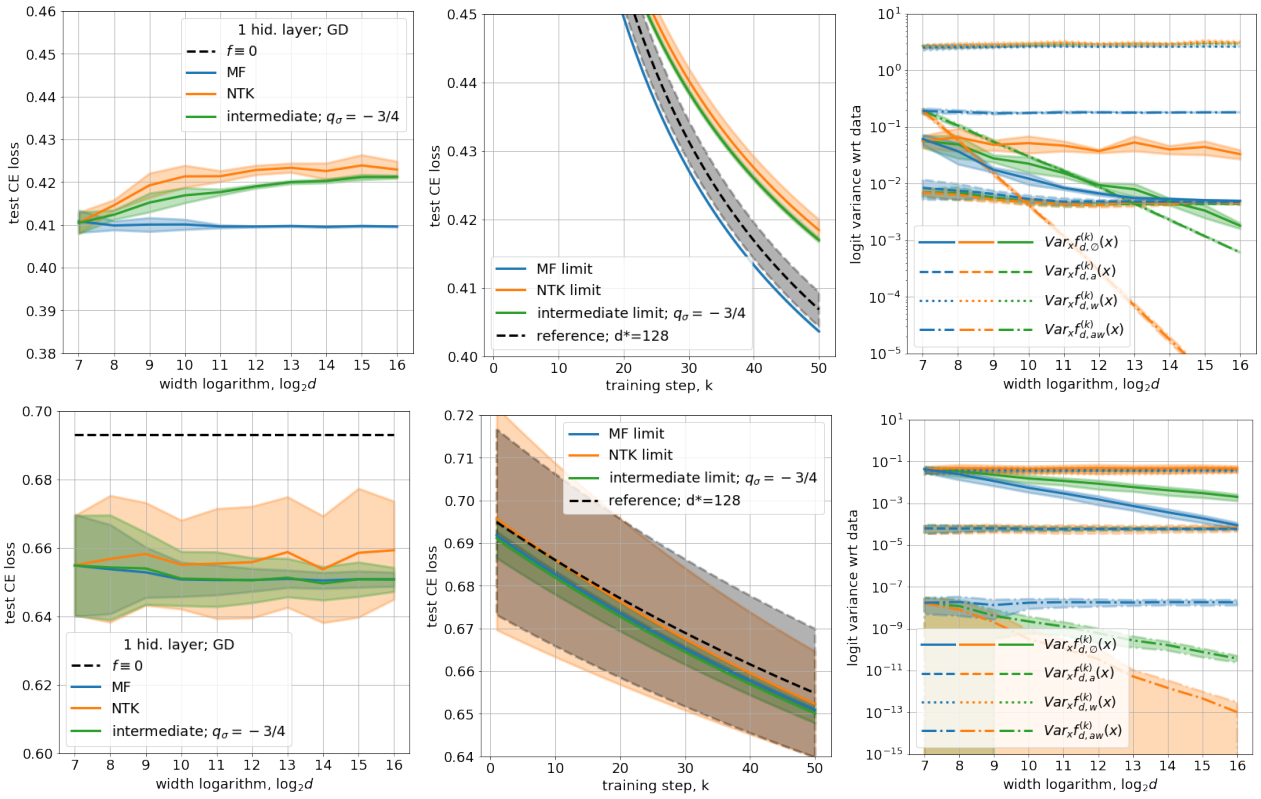

Figure 3. **For small learning rates, the NTK limit approximates the reference finite-width network better than the MF limit.** *Top row:* scaling a reference network trained with gradient descent with (unscaled) learning rates $\eta_a^* = \eta_w^* = 0.02$. *Bottom row:* same with unscaled learning rates $\eta_a^* = \eta_w^* = 0.0002$. *Left:* a final test cross-entropy (CE) loss as a function of width d. *Center:* test CE loss as a function of training step k for a reference net and its limits. As one can see, MF and intermediate limits preserve mean CE loss but not its variance with respect to the initialization. In contrast, the NTK limit does preserve the variance. *Right:* variance with respect to the data distribution for terms of model decomposition (6) as a function of width d. When learning rates are small, $f^{(k)}_{d,\emptyset}$, which contributes to the variance, becomes the largest term in decomposition (6) and $f^{(k)}_{d,aw}$, which vanishes in NTK and intermediate limits, becomes the smallest. As we have noticed in Figure 2 for the MF limit $f^{(k)}_{d,\emptyset}$ is not exactly constant but decays approaching a constant for large d. This is the reason for the MF limit not to preserve the variance of CE loss. *Setup:* We train a 1-hidden layer net on a subset of CIFAR2 (a dataset of first two classes of CIFAR10) of size 1000 with gradient descent. We take a reference net of width $d^* = 2^7 = 128$ and scale its hyperparameters according to MF (blue curves), NTK (orange curves) and intermediate scalings with $q_\sigma = -3/4$ (green curves, see text). See SM A for further details.

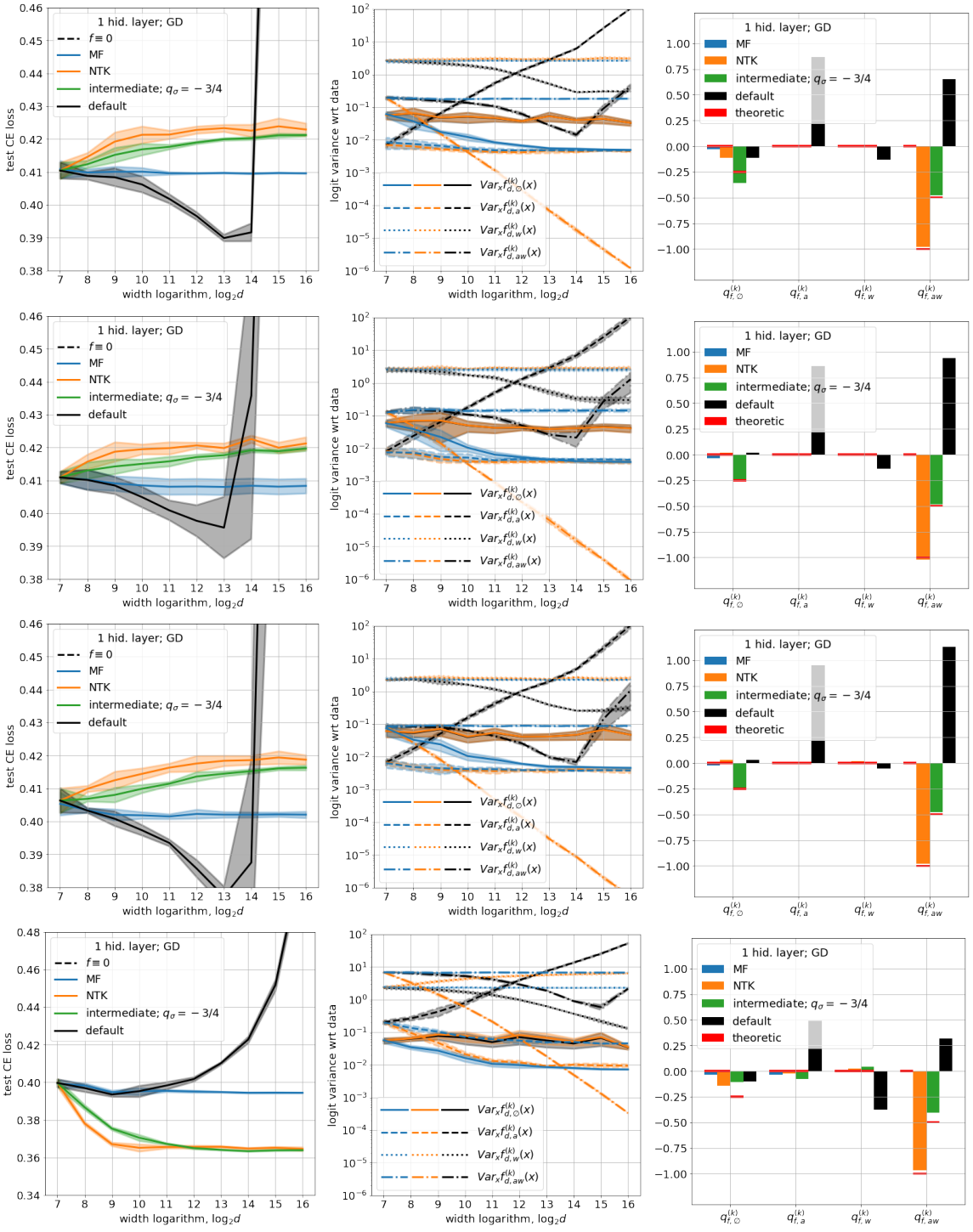

Figure 4. **Applying a mini-batch instead of a full batch gradient descent does not introduce any qualitative changes. The same holds for training on a larger dataset.** *Top row:* scaling a reference network trained with a full-batch GD with (unscaled) learning rates $\eta_a^* = \eta_w^* = 0.02$ for 50 gradient steps on a subset of CIFAR2 (a dataset of first two classes of CIFAR10) of size 1000. *Second row:* same with a mini-batch GD with batches of size 100. *Third row:* same as the top row but on a full CIFAR2 (10000 training samples) with the mini-batch GD with batches of size 1000. *Bottom row:* same as the top row but with 500 gradient steps. *Left:* a final test cross-entropy (CE) loss as a function of width d. *Center:* variance with respect to the data distribution for terms of model decomposition (6) as a function of width d. *Right:* numerical estimates for exponents of decomposition (6) terms, as well as their theoretical values (denoted by red ticks). See SM A for further details.

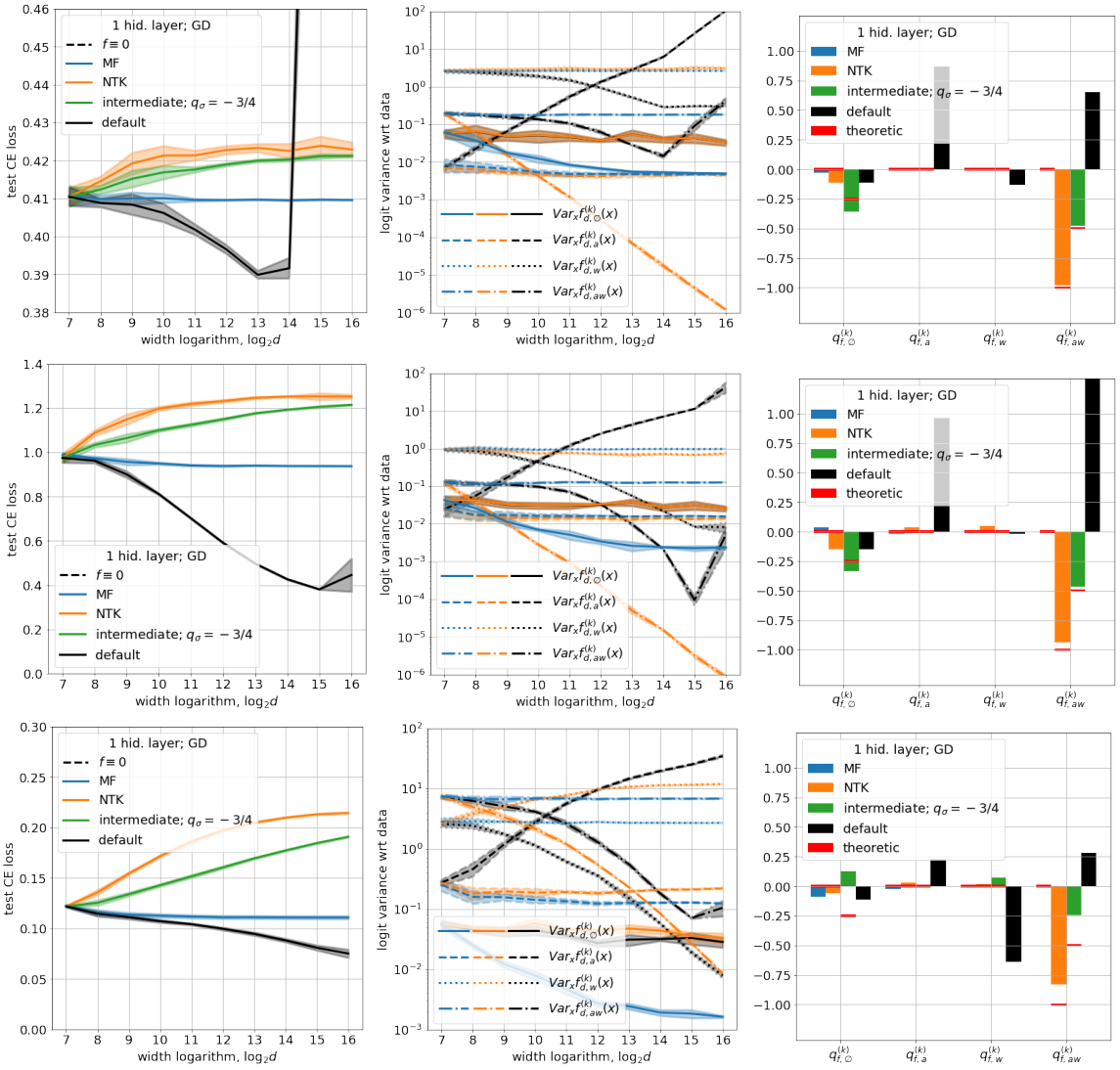

Figure 5. **Considering a multi-class classification instead of a binary one does not introduce any qualitative changes.** *Top row:* scaling a reference network trained with a full-batch GD with (unscaled) learning rates $\eta_a^* = \eta_w^* = 0.02$ for 50 gradient steps on a subset of CIFAR2 (a dataset of first two classes of CIFAR10) of size 1000. *Middle row:* same for a subset of MNIST of size 1000. *Bottom row:* scaling a reference network trained with SGD using batches of size 100 with (unscaled) learning rates $\eta_a^* = \eta_w^* = 0.02$ for 6000 gradient steps on MNIST dataset. *Left:* a final test cross-entropy (CE) loss as a function of width d. *Center:* variance with respect to the data distribution for terms of model decomposition (6) as a function of width d. *Right:* numerical estimates for exponents of decomposition (6) terms, as well as their theoretical values (denoted by red ticks). See SM A for further details.