A Appendix

A.1 Proof of Proposition 2.3

Lemma A.1. *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of random variables with infinite mean and variance. Let us assume that there exists a random variable* X *with finite mean and variance such that* $X_n \xrightarrow{d} X$, *where* d *denotes convergence in distribution. Given the number of samples* K, *let the empirical mean and variance random variables* $(\hat{\mu}_n)_{n\in\mathbb{N}}, (\hat{\sigma}_n)_{n\in\mathbb{N}}$ *defined as*

$$\begin{aligned}\forall n \in \mathbb{N}, \quad \hat{\mu}_n &= \frac{1}{K}\sum_{k=1}^{K} X_{n,k} \\ \hat{\sigma}_n &= \frac{1}{K-1}\sum_{k=1}^{K}(X_{n,k} - \hat{\mu}_n)^2\end{aligned}$$

where the $X_{n,k}$ *are i.i.d samples. Then we have*

$$\hat{\mu}_n \xrightarrow{d} \hat{\mu} \quad \text{as well as} \quad \hat{\sigma}_n \xrightarrow{d} \hat{\sigma} \tag{12}$$

where $\hat{\mu}, \hat{\sigma}$ *are resp. the empirical mean and variance of the limiting distribution* X.

The proof is a straightforward application of the continuous mapping theorem.

Lemma A.2. *Let* $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ *be a sequence of random variables such that* $|X_n| = \mathcal{O}_p(1)$ *and* $|Y_n| = \mathcal{O}_p(1)$. *Let us assume that there exists a real valued sequence* $(a_n)_{n\in\mathbb{N}}$ *such that* $|X_n - Y_n| = \mathcal{O}_p(a_n)$. *Given* K, *let the empirical variance random variables* $(\hat{\sigma}_n^X)_{n\in\mathbb{N}}$ *defined as*

$$\forall n \in \mathbb{N}, \quad \hat{\sigma}_n^X = \frac{1}{K-1}\sum_{k=1}^{K}(X_{n,k} - \frac{1}{K}\sum_{k=1}^{K} X_{n,k})^2$$

We define similarly $(\hat{\sigma}_n^Y)_{n\in\mathbb{N}}$. *Then,*

$$|\hat{\sigma}_n^X - \hat{\sigma}_n^Y| = \mathcal{O}_p(a_n)$$

Proof.

$$\begin{aligned}|\hat{\sigma}_n^X - \hat{\sigma}_n^Y| =& \Big|\frac{1}{K-1}\sum_{k'=1}^{K}\Big[(X_{n,k'} - \frac{1}{K}\sum_{k=1}^{K} X_{n,k})^2 - (Y_{n,k'} - \frac{1}{K}\sum_{k=1}^{K} Y_{n,k})^2\Big]\Big| \\ =& \frac{1}{K-1}\Big|\sum_{k'=1}^{K}\Big[(X_{n,k'} - \frac{1}{K}\sum_{k=1}^{K} X_{n,k} - Y_{n,k'} + \frac{1}{K}\sum_{k=1}^{K} Y_{n,k})(X_{n,k'} - \frac{1}{K}\sum_{k=1}^{K} X_{n,k} + Y_{n,k'} - \frac{1}{K}\sum_{k=1}^{K} Y_{n,k})\Big]\Big| \\ =& \frac{1}{K-1}\Big|\sum_{k'=1}^{K}\Big[(X_{n,k'} - Y_{n,k'} - \frac{1}{K}\sum_{k=1}^{K}(X_{n,k} - Y_{n,k}))(X_{n,k'} + Y_{n,k'} - \frac{1}{K}\sum_{k=1}^{K}(X_{n,k} + Y_{n,k}))\Big]\Big| \\ =& \mathcal{O}_p(\Big|\sum_{k'=1}^{K}\Big[(X_{n,k'} - Y_{n,k'} - \frac{1}{K}\sum_{k=1}^{K}(X_{n,k} - Y_{n,k}))\Big]\Big|) \\ =& \mathcal{O}_p(\sum_{k=1}^{K}|X_{n,k} - Y_{n,k}|) \\ =& \mathcal{O}_p(a_n)\end{aligned}$$

□

A.1.1 Discrepancy between non-linear and linearly trained neural network during training

We adapt Theorem H.1. from [25] to show that the discrepancy between the original and the linearly trained network for the MSE loss that we consider is bounded as $\sup_t \|f_t^{lin}(x) - f(x, \theta_t)\|_2^2 \leq \mathcal{O}(h^{-2})$ as well. The proof is an adaptation of the one in [25] with very minor differences. We piggyback on the main result in their proof which was obtained with Grönwall's inequality, which requires the continuity of the derivative of the activation function.

Let a neural network f such the width of all hidden layers are identical, $h_1 = h_2 = ... = h_{L-1} = h$, and such that ϕ' is bounded and Lipschitz continuous on $\mathbb{R}$. Let the training data $(\mathcal{X}, \mathcal{Y})$ contained in some compact set, such that the NTK of f on $\mathcal{X}$ is invertible. Let f_t the model trained on the MSE loss with gradient flow at timestep t with some learning rate.

Assumption A.3. $\forall \delta > 0, \exists C, N : \forall h > H$, with probability at least $1 - \delta$ over random initialization,

$$\sup_t \|\Theta_{\theta_0} - \Theta_{\theta_t}\|_F \leq \frac{C}{h} \tag{13}$$

Proposition A.4. *Under assumption A.3, when trained with gradient flow on the MSE loss, we have* $\forall x, \forall \delta > 0, \exists C, H : \forall h > H,$

$$\mathbb{P}\big[\sup_t \|f_t^{lin}(x) - f_t(x)\|_2 \leq \frac{C}{h}\big] \geq 1 - \delta \tag{14}$$

Proof. Let $g^{lin}(t) = f_t^{lin}(\mathcal{X}) - \mathcal{Y}$ and $g(t) = f_t(\mathcal{X}) - \mathcal{Y}$. Starting from equation (S118) from [25], we have

$$\|g^{lin}(t) - g(t)\|_2 \leq \eta_0 t \sigma_t e^{-\lambda_0 \eta_0 t + \sigma_t \eta_0 t} \|g(0)\|_2 \tag{15}$$

where $\sigma_t = \sup_{0 \leq s \leq t} \|\Theta_{\theta_t} - \Theta_{\theta_0}\|_{op}$, η_0 the learning rate and λ_0 is the smallest eigenvalue of Θ_{θ_0}.

Because the functions are trained on MSE loss L, we have

$$\begin{aligned}
\frac{d}{dt}(f_t^{lin}(x) - f_t(x)) =& \eta_0 \Theta_{\theta_0}(x, \mathcal{X}) L'(f_t^{lin}) - \eta_0 \Theta_{\theta_t}(x, \mathcal{X}) L'(f_t) && (16)\\
=& \eta_0 \Theta_{\theta_0}(x, \mathcal{X}) g^{lin}(t) - \eta_0 \Theta_{\theta_t}(x, \mathcal{X}) g(t) && (17)\\
=& \eta_0 (\Theta_{\theta_0}(x, \mathcal{X}) - \Theta_{\theta_t}(x, \mathcal{X})) g^{lin}(t) - \eta_0 \Theta_{\theta_t}(x, \mathcal{X})(g(t) - g^{lin}(t)) && (18)
\end{aligned}$$

Integrating and taking the L2 norm,

$$\begin{aligned}
\|f_t^{lin}(x) - f_t(x))\| \leq& \eta_0 \int_0^t \|(\Theta_{\theta_0}(x, \mathcal{X}) - \Theta_{\theta_{t'}}(x, \mathcal{X}))\| \|g^{lin}(t')\| dt' && (19)\\
&+ \eta_0 \int_0^t \|\Theta_{\theta_{t'}}(x, \mathcal{X})\| \|g(t') - g^{lin}(t')\| dt' && (20)\\
\leq& \eta_0 \|g(0)\| \int_0^t \|(\Theta_{\theta_0}(x, \mathcal{X}) - \Theta_{\theta_{t'}}(x, \mathcal{X}))\| e^{-\lambda_0 \eta_0 t'} dt' && (21)\\
&+ \eta_0 \int_0^t [\|\Theta_{\theta_0}(x, \mathcal{X})\| + \|\Theta_{\theta_{t'}}(x, \mathcal{X}) - \Theta_{\theta_0}(x, \mathcal{X})\| && (22)\\
&\cdot \|g(0)\| \eta_0 t' \sigma_{t'} e^{-\lambda_0 \eta_0 t' + \sigma_{t'} \eta_0 t'} dt' && (23)\\
& && (24)
\end{aligned}$$

where we used $\|g(0)\| = \|g^{lin}(t')\| \leq \|g^{lin}(0)\| e^{-\lambda_0 \eta_0 t'}$ the triangular inequality and equation 15.

Because $g(0)$ converges in distribution to a mean zero gaussian distribution, and because Θ_{θ_0} converges in probability to Θ_∞, we can find H such that $\forall h > H$, with probability at least $1 - \delta'$,

$$\|g(0)\|_2 \leq C \tag{25}$$

and

$$\|\Theta_{\theta_0}(x, \mathcal{X})\|_2 \leq C \tag{26}$$

where $C > 0$ is a constant.

Because the NTK at initialization converges in probaility to Θ_∞ assumed to be invertible, there exists H' such that $\forall h > H'$,

$$\|\Theta_{\theta_0} - \Theta_\infty\|_F \leq \frac{\lambda_{min}}{2} \tag{27}$$

Where λ_{min} is the smallest eigenvalue of Θ_∞. Thus $\|\Theta_{\theta_0} - \Theta_\infty\|_{op} \leq \frac{\lambda_{min}}{2}$, and so $\lambda_0 > \frac{\lambda_{min}}{2}$

From assumption A.3, let us fix H'', C' such that $\forall h > H''$, with probability at least $1 - \delta'$

$$\sigma_t = \sup_{0 \leq t' \leq t} \|\Theta_{\theta'_t} - \Theta_{\theta_0}\|_{op} \leq \sup_{0 \leq s \leq t} \|\Theta_{\theta_t} - \Theta_{\theta_0}\|_F \leq \frac{C'}{h} \tag{28}$$

and

$$\sup_{0 \leq t' \leq t} \|\Theta_{\theta'_t}(x, \mathcal{X}) - \Theta_{\theta_0}(x, \mathcal{X})\|_2 \leq \frac{C'}{h} \tag{29}$$

And therefore $\forall h > \max(H', H'', \frac{2C'}{\lambda min})$, with probability at least $1 - \delta'$, $\sigma_t < \lambda_0$, and therefore $\int_0^t t' e^{-\lambda_0 \eta_0 t' + \sigma_{t'} \eta_0 t'}$ is bounded by some C''.

Putting everything together, $\forall n > \max(H, H', H'', \frac{2C'}{\lambda min})$, with probability at least $1 - 3\delta'$,

$$\|f_t^{lin}(x) - f_t(x))\| \leq \eta_0 C \int_0^t \frac{C'}{h} e^{-\lambda_{min}\eta_0 t'} dt' + \eta_0 [C + \frac{C'}{h}] C \eta_0 \frac{C'}{h} C'' \leq \frac{K}{h} \tag{30}$$

$$\tag{31}$$

with K some constant. By taking $\delta' = \frac{\delta}{3}$ we get the result that $\|f_t^{lin}(x) - f_t(x))\| = \mathcal{O}_p(\frac{1}{h})$

Finally, using Lemma A.2 and the fact that $f^{lin}(x)$ and $f(x)$ are bounded with high probability since they both converge in distribution to a gaussian with finite variance, we have, at the end of training,

$$|\hat{\mathbb{V}}(f(x)) - \hat{\mathbb{V}}(f^{\text{lin}}(x))| = \mathcal{O}_p(\frac{1}{h})$$

for some finite sample empirical variance.

It remains to show that $\forall x, \forall \delta, \exists C > 0, H > 0 : \forall h > H$,

$$\mathbb{P}\big[\frac{1}{h} \leq C \hat{\mathbb{V}}\big[(\mathcal{Q}_{\theta_0}(x, \mathcal{X}) - \bar{\mathcal{Q}}(x, \mathcal{X}))(\mathcal{Y} - f(\mathcal{X}, \theta_0))\big]\big] \geq 1 - \delta \tag{32}$$

i.e.

$$\mathbb{P}\big[\frac{1}{C} \leq \hat{\mathbb{V}}\big[\sqrt{h}(\mathcal{Q}_{\theta_0}(x, \mathcal{X}) - \bar{\mathcal{Q}}(x, \mathcal{X}))(\mathcal{Y} - f(\mathcal{X}, \theta_0))\big]\big] \geq 1 - \delta \tag{33}$$

Following Proposition 2.1 and Lemma A.1, we have

$$\hat{\mathbb{V}}\big[\sqrt{h}(\mathcal{Q}_{\theta_0}(x, \mathcal{X}) - \bar{\mathcal{Q}}(x, \mathcal{X}))(\mathcal{Y} - f(\mathcal{X}, \theta_0))\big] \xrightarrow{d} \hat{\mathbb{V}}\big[Z(x)\big]$$

where Z is a linear combination of 2 chi-square distribution with finite and no-zero moments, which proves the result.

□

A.2 Delta method

A.2.1 Proof of Proposition 2.1

We start with the special case of a single hidden layer neural network. We provide the following Lemma, which is a slight variation of the Delta method.

Lemma A.5. *Let $X_h \in \mathbb{R}^n, Y_h \in \mathbb{R}^n$ be two sequences of multivariate independent random variables that satisfy $X_h \overset{dist.}{\to} \mathcal{N}(\mu, \Sigma_1)$ and $\sqrt{h}(Y_h - \bar{Y}) \overset{dist.}{\to} \mathcal{N}(0, \Sigma_2)$ in distribution for some constant $\bar{Y}$. Let a function $g : \mathbb{R}^n \to \mathbb{R}^n$ with continuous partial derivative. Then,*

$$\sqrt{h}\big[g(Y_h)^T X_h - g(\bar{Y})^T X_h\big] \overset{dist.}{\to} Z \tag{34}$$

such that Z is a linear combination of 2 Chi-square distributions, and

$$\mathbb{E}[Z] = 0 \tag{35}$$

$$\mathbb{V}[Z] = Tr(\nabla^T g(\bar{Y})\Sigma_2 \nabla g(\bar{Y})\Sigma_1) + Tr(\nabla^T g(\bar{Y})\Sigma_2 \nabla g(\bar{Y})\mu\mu^T) \tag{36}$$

Proof. By applying the multivariate delta method, we have

$$\sqrt{h}\big[g(Y_h) - g(\bar{Y})\big] \overset{dist.}{\to} \mathcal{N}(0, \nabla^T g(\bar{Y})\Sigma_2 \nabla g(\bar{Y})) \tag{37}$$

Given the independence assumption of X_h and Y_h, we have the independence of X_h and $\sqrt{h}\big[g(Y_h)^T - g(\bar{Y})^T\big]$, and therefore $(X_h, \sqrt{h}\big[g(Y_h)^T - g(\bar{Y})^T\big])$ converge in distribution to the Cartesian product of their respective limiting random variables. Using the continuity of the dot-product operation, and applying again the continuous mapping theorem, we have

$$\sqrt{h}\big[g(Y_h) - g(\bar{Y})\big]^T X_h \overset{dist.}{\to} Z = \mathcal{G}_1^T \mathcal{G}_2 \tag{38}$$

where $\mathcal{G}_1, \mathcal{G}_2$ are normally distributed multivariate random variables with (mean, covariance) resp. $(0, \nabla^T g(\bar{Y})\Sigma_2 \nabla g(\bar{Y}))$ and (μ, Σ_1).

Note that if the X_h are constant, or converge to a constant value, the limiting distribution Z is a Gaussian distribution. In general however, given the independence of $\mathcal{G}_1$ and $\mathcal{G}_2$, Z as the product of 2 independent Gaussians is a linear combination of two Chi-square distributions.

Finally, we have

$$\mathbb{E}[Z] = 0 \tag{39}$$

$$\mathbb{V}[Z] = \mathbb{E}[\mathcal{G}_2^T \mathcal{G}_1 \mathcal{G}_1^T \mathcal{G}_2] \tag{40}$$

$$= \mathbb{E}[Tr(\mathcal{G}_2^T \mathcal{G}_1 \mathcal{G}_1^T \mathcal{G}_2)] \tag{41}$$

$$= \mathbb{E}[Tr(\mathcal{G}_1 \mathcal{G}_1^T \mathcal{G}_2 \mathcal{G}_2^T)] \tag{42}$$

$$= Tr(\mathbb{E}[\mathcal{G}_1 \mathcal{G}_1^T \mathcal{G}_2 \mathcal{G}_2^T]) \tag{43}$$

$$= Tr(\mathbb{E}[\mathcal{G}_1 \mathcal{G}_1^T]\mathbb{E}[\mathcal{G}_2 \mathcal{G}_2^T]) \tag{44}$$

$$= Tr(\nabla^T g(\bar{Y})\Sigma_2 \nabla g(\bar{Y})[\Sigma_1 + \mu\mu^T]) \tag{45}$$

which concludes the lemma.

□

Let us now prove Proposition 2.1. For one hidden layer networks, given a width h, it is straightforward to see (see A.4) that the empirical NTK Θ_h (whereby the weight initialization is a random variable) is the sum of h i.i.d. random variables which mean equals the infinite width NTK Θ_∞, i.e

$\forall(\mathcal{X}, \mathcal{X}')$,

$$\Theta_h(\mathcal{X}, \mathcal{X}') = \frac{1}{h}\sum_i \hat{\Theta}^i(\mathcal{X}, \mathcal{X}') \tag{46}$$

$$\hat{\Theta}^i(\mathcal{X}, \mathcal{X}') \sim_{i.i.d} \hat{\Theta}(\mathcal{X}, \mathcal{X}') \tag{47}$$

$$\mathbb{E}[\hat{\Theta}(\mathcal{X}, \mathcal{X}')] = \Theta_\infty(\mathcal{X}, \mathcal{X}') \tag{48}$$

Proposition A.6. *For one hidden layer networks,*

$$\sqrt{h}[\mathcal{Q}_{\theta_0}(x, \mathcal{X}) - \bar{\mathcal{Q}}(x, \mathcal{X})](\mathcal{Y} - f(\mathcal{X}, \theta_0)) \overset{d}{\to} Z$$

where Z is the linear combination of 2 Chi-Square distributions, and

$$
\begin{aligned}
\mathbb{E}[Z] =& 0 \\
\mathbb{V}[Z] =& \mathbb{V}_1^c(x) + \mathbb{V}_1^i(x) \\
\mathbb{V}_1^c(x) =& \mathbb{V}[\hat{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}] + \mathbb{V}[\bar{\mathcal{Q}}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}] \\
& - 2\mathbb{C}ov[\bar{\mathcal{Q}}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}, \hat{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}], \\
\mathbb{V}_1^i(x) =& \mathbb{V}[\hat{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] + \mathbb{V}[\bar{\mathcal{Q}}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] \\
& - 2\mathbb{C}ov[\bar{\mathcal{Q}}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0), \hat{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)].
\end{aligned}
$$

Proof. Following the Central Limit Theorem, we have the following convergence in distribution:

$$\sqrt{h}\big[\Theta_h(\mathcal{X},\mathcal{X}') - \Theta_\infty(\mathcal{X},\mathcal{X}')\big] \overset{dist.}{\rightarrow} \mathcal{N}(0,\Sigma) \tag{49}$$

where Σ is the covariance matrix between the entries of $\hat{\Theta}(\mathcal{X},\mathcal{X}')$.

Let the function

$$g(W,v) = v^T W^{-1} \tag{50}$$

for any invertible block matrix W, and vector v.

Note that $g(\Theta_h(\mathcal{X},\mathcal{X}),\Theta_h(\mathcal{X},x))(\mathcal{Y} - f(\mathcal{X},\theta_0))$ is the prediction of a linearly trained neural network evaluated on x trained on $\mathcal{X}$, given a functional initialization $f(.,\theta_0)$ and NTK Θ_h. We wish to estimate the asymptotic behavior of the expectation and variance of this quantity in the limit of $h \to \infty$. However, these moments are not always defined because the support of $\Theta_h(\mathcal{X},\mathcal{X})$ contains non invertible instances of the gram schmidt matrix (e.g. all weights initialized at 0), which induces divergent moments. However, because of the convergence in probability of Θ_h to Θ_∞ (which is invertible by assumption), the event of such singularities becomes rarer as h increases, and the delta method allows us to get the asymptotic expectation and variance.

Using the fact that g has continuous first partial derivatives, and the independence of $f(\mathcal{X},\theta_0)$ and Θ_h, following Lemma A.5,

$$\sqrt{h}\big[g(\Theta_h(\mathcal{X},\mathcal{X}),\Theta_h(\mathcal{X},x))f(\mathcal{X},\theta_0) - g(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x))f(\mathcal{X},\theta_0)\big] \overset{dist.}{\rightarrow} Z \tag{51}$$

with Z being the linear combination of 2 Chi-Square distributions, and

$$\mathbb{E}[Z] = 0 \tag{52}$$

$$\mathbb{V}[Z] = Tr\Big(\Sigma\mathcal{K}(\mathcal{X},\mathcal{X})\Big) + Tr\Big(\Sigma\mathcal{Y}\mathcal{Y}^T\Big) \tag{53}$$

$$\tag{54}$$

where, by vectorizing matrices and using g_k the k-th entry of the value of g,

$$
\begin{aligned}
\Sigma_{i,j} =& \nabla_W^T g_i(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x))\mathbb{C}ov[vect(\hat{\Theta}(\mathcal{X},\mathcal{X}))]\nabla_W g_j(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x)) \\
&+ \nabla_v^T g_i(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x))\mathbb{C}ov[\hat{\Theta}(\mathcal{X},x)]\nabla_v g_j(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x)) \\
&+ \nabla_W^T g_i(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x))\mathbb{C}ov[vect(\hat{\Theta}(\mathcal{X},\mathcal{X})),\hat{\Theta}(\mathcal{X},x)]\nabla_v g_j(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x)) \\
&+ \nabla_v g_i(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x))\mathbb{C}ov[vect(\hat{\Theta}(\mathcal{X},\mathcal{X})),\hat{\Theta}(\mathcal{X},x)]\nabla_W^T g_j(\Theta_\infty(\mathcal{X},\mathcal{X}),\Theta_\infty(\mathcal{X},x))
\end{aligned}
$$

Using $\nabla_v g_k(W,v) = W^{-1}u_k$ and $\nabla_W g_k = vect(-W^{-T}vu_k^TW^{-T})$, where u_k is the vector 0 everywhere except for the k-th position which is 1, the expression can be rewritten as

$$
\begin{aligned}
\Sigma =& \mathbb{C}ov[\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\hat{\Theta}(\mathcal{X},\mathcal{X})^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\Theta_\infty(\mathcal{X},x)] \\
&+ \mathbb{C}ov[\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\hat{\Theta}(\mathcal{X},x)] \\
&- 2\mathbb{C}ov[\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\hat{\Theta}(\mathcal{X},\mathcal{X})^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\Theta_\infty(\mathcal{X},x), \Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\hat{\Theta}(\mathcal{X},x)]
\end{aligned}
$$

Finally, we can notice that the following expression equals that of 53

$$
\begin{aligned}
&\mathbb{V}[\hat{\Theta}(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] \\
&+\mathbb{C}ov[\Theta_\infty(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\hat{\Theta}(\mathcal{X},\mathcal{X})^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] \\
&-2\Theta_\infty(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\mathbb{C}ov[\hat{\Theta}(\mathcal{X},\mathcal{X})^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0),\hat{\Theta}(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] \\
&+\mathbb{V}[\hat{\Theta}(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}] \\
&+\mathbb{C}ov[\Theta_\infty(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\hat{\Theta}(\mathcal{X},\mathcal{X})^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}] \\
&-2\Theta_\infty(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\mathbb{C}ov[\hat{\Theta}(\mathcal{X},\mathcal{X})^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y},\hat{\Theta}(\mathcal{X},x)^T\Theta_\infty(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}]
\end{aligned}
$$

which concludes the proof by using $\mathbb{E}[\hat{\Theta}]=\Theta_\infty$. □

A.2.2 Approximation in the general case

In the general case, we can no longer apply the central limit theorem to asymptotically describe the NTK as a gaussian. Nonetheless, the delta method is often used in a form that is essentially identical to that above, but without the asymptotically normal assumption, so long as the fluctuation of the variable around the mean vanishes, i.e. $\|\Theta_{\theta_0}(\{\mathcal{X},x\},\{\mathcal{X},x\})-\bar{\Theta}(\{\mathcal{X},x\},\{\mathcal{X},x\})\|_F=o_p(1)$.

Using the identity $M^{-1}=\bar{M}^{-1}+\bar{M}^{-1}(\bar{M}-M)\bar{M}^{-1}+\bar{M}^{-1}\big[(\bar{M}-M)\bar{M}^{-1}\big]^2\bar{M}M^{-1}$ for any pair of invertible matrices $M,\bar{M}$, we can rewrite f^{lin} as

$$
\begin{aligned}
f^{\text{lin}}(x)=&f(x)+\Theta_{\theta_0}(x)\Theta_{\theta_0}^{-1}(\mathcal{Y}-f) \\
=&f(x)+\Theta_{\theta_0}(x)[\bar{\Theta}^{-1}+\bar{\Theta}^{-1}(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}+\bar{\Theta}^{-1}\big[(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}\big]^2\bar{\Theta}\Theta_{\theta_0}^{-1}](\mathcal{Y}-f) \\
=&f(x)+\bar{\Theta}(x)[\bar{\Theta}^{-1}+\bar{\Theta}^{-1}(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}+\bar{\Theta}^{-1}\big[(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}\big]^2\bar{\Theta}\Theta_{\theta_0}^{-1}](\mathcal{Y}-f) \\
&+[\Theta_{\theta_0}(x)-\bar{\Theta}(x)][\bar{\Theta}^{-1}+\bar{\Theta}^{-1}(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}+\bar{\Theta}^{-1}\big[(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}\big]^2\bar{\Theta}\Theta_{\theta_0}^{-1}](\mathcal{Y}-f) \\
=&f(x)+\bar{\Theta}(x)\bar{\Theta}^{-1}(\mathcal{Y}-f) \\
&+\bar{\Theta}(x)\bar{\Theta}^{-1}(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}(\mathcal{Y}-f)+(\Theta_{\theta_0}(x)-\bar{\Theta}(x))\bar{\Theta}^{-1}(\mathcal{Y}-f) \\
&+\big[\Theta_{\theta_0}(x)-\bar{\Theta}(x)\bar{\Theta}^{-1}\Theta_{\theta_0}\big]\bar{\Theta}^{-1}(\bar{\Theta}-\Theta_{\theta_0})\bar{\Theta}^{-1}(\mathcal{Y}-f) \\
&+o_p(\|\Theta_{\theta_0}(\{\mathcal{X},x\},\{\mathcal{X},x\})-\bar{\Theta}(\{\mathcal{X},x\},\{\mathcal{X},x\})\|_F^2)
\end{aligned}
$$

where we note $\Theta_{\theta_0}=\Theta_{\theta_0}(\mathcal{X},\mathcal{X}),\Theta_{\theta_0}(x)=\Theta_{\theta_0}(\mathcal{X},x)$ (resp. for $\bar{\Theta},f$) for ease of notation.

For sufficiently large width, with high probability the remainder term will be negligible. Keeping the empirical mean and variance in mind, we can now take the expectation and variance ignoring the rare singularities.

$$
\mathbb{E}[f^{\text{lin}}(x)]\approx\bar{\Theta}(x)\bar{\Theta}^{-1}\mathcal{Y}+\mathbb{E}\big[[\Theta_{\theta_0}(x)-\bar{\Theta}(x)\bar{\Theta}^{-1}\Theta_{\theta_0}][\bar{\Theta}^{-1}(\bar{\Theta}-\Theta_{\theta_0})]\big]\bar{\Theta}^{-1}\mathcal{Y}
$$

$$
\begin{aligned}
\mathbb{V}[f^{\text{lin}}(x)] \approx & \bar{\mathcal{K}}(x,x) + \bar{\Theta}(x)\bar{\Theta}^{-1}\bar{\mathcal{K}}(\mathcal{X},\mathcal{X})\bar{\Theta}^{-1}\bar{\Theta}(x)^T - 2\bar{\Theta}(x)\bar{\Theta}^{-1}\bar{\mathcal{K}}(\mathcal{X},x) \\
& + \mathbb{V}[(\Theta_{\theta_0}(x) - \bar{\Theta}(x))\bar{\Theta}^{-1}(\mathcal{Y} - f)] \\
& + \mathbb{V}[\bar{\Theta}(x)\bar{\Theta}^{-1}(\bar{\Theta} - \Theta_{\theta_0})\bar{\Theta}^{-1}(\mathcal{Y} - f)] \\
& + 2\mathbb{C}ov[\bar{\Theta}(x)\bar{\Theta}^{-1}(\bar{\Theta} - \Theta_{\theta_0})\bar{\Theta}^{-1}(\mathcal{Y} - f), (\Theta_{\theta_0}(x) - \bar{\Theta}(x))\bar{\Theta}^{-1}(\mathcal{Y} - f)] \\
& - 2\mathbb{C}ov[f(x) - \bar{\Theta}(x)\bar{\Theta}^{-1}f, \big[\Theta_{\theta_0}(x) - \bar{\Theta}(x)\bar{\Theta}^{-1}\Theta_{\theta_0}\big]\bar{\Theta}^{-1}(\bar{\Theta} - \Theta_{\theta_0})\bar{\Theta}^{-1}f] \\
= & \bar{\mathcal{K}}(x,x) + \bar{\mathcal{Q}}(x,\mathcal{X})\bar{\mathcal{K}}(\mathcal{X},\mathcal{X})\bar{\mathcal{Q}}(x,\mathcal{X})^T - 2\bar{\mathcal{Q}}(x,\mathcal{X})\bar{\mathcal{K}}(\mathcal{X},x) \\
& + \mathbb{V}[\Theta_{\theta_0}(x)\bar{\Theta}^{-1}(\mathcal{Y} - f)] \\
& + \mathbb{V}[\bar{\mathcal{Q}}(x,\mathcal{X})\Theta_{\theta_0}\bar{\Theta}^{-1}(\mathcal{Y} - f)] \\
& - 2\mathbb{C}ov[\bar{\mathcal{Q}}(x,\mathcal{X})\Theta_{\theta_0}\bar{\Theta}^{-1}(\mathcal{Y} - f), \Theta_{\theta_0}(x)\bar{\Theta}^{-1}(\mathcal{Y} - f)] \\
& - 2\mathbb{E}\big[[\Theta_{\theta_0}(x) - \bar{\mathcal{Q}}(x,\mathcal{X})\Theta_{\theta_0}][\bar{\Theta}^{-1}(\bar{\Theta} - \Theta_{\theta_0})\bar{\Theta}^{-1}]\big]\big[\bar{\mathcal{K}}(\mathcal{X},x) - \bar{\mathcal{K}}(\mathcal{X},\mathcal{X})\bar{\mathcal{Q}}(x,\mathcal{X})^T\big] \\
= & \mathbb{V}^a(x) + \mathbb{V}^c(x) + \mathbb{V}^i(x) + \mathbb{V}^{cor}(x)
\end{aligned}
$$

Assuming the fluctuation of Θ_{θ_0} around its mean is in the order of $\mathcal{O}(h^{-\frac{1}{4}})$ (see A.2.3), we have $\mathbb{V}^c(x), \mathbb{V}^i(x), \mathbb{V}^{cor}(x)$ all of order $\mathcal{O}(\frac{1}{h})$. While the variance of the true residual might not be finite for the same reason as why $\mathbb{V}[f(x)]$ is not, expanding the approximation to one order higher yields $\mathbb{V}^{res}(x) \approx \mathcal{O}(\frac{1}{h^2})$.

A.2.3 Fluctuation of the NTK initialization

For one hidden layer networks, we can bound the fluctuation of the NTK at initialization using the Central Limit Theorem, which yields $\mathbb{V}[\Theta_{\theta_0}] = \mathcal{O}(\frac{1}{\sqrt{h}})$.

While we do not provide a proof, a similar heuristic argument presented in Appendix C. of [38] can be used to argue for the same bound in the general case of arbitrary depth networks.

A.3 Other remarks

A.3.1 Interpretation of variance terms

Given a centered functional initialization g and an NTK $\bar{\Theta}$, a fully trained neural network has the functional expression

$$f(x) = g(x) + \bar{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}(\mathcal{Y} - g(\mathcal{X})). \tag{55}$$

The predictive variance is then

$$
\begin{aligned}
\mathbb{V}[f(x)] = & \mathbb{E}[(g(x) - \bar{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}g(\mathcal{X}))^2] && (56) \\
= & \mathbb{E}[g(x)^2 + \bar{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}g(\mathcal{X})g(\mathcal{X})^T\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\bar{\Theta}(x,\mathcal{X})^T && (57) \\
& - 2\bar{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}g(\mathcal{X})g(x)] && (58) \\
= & \mathbb{V}[g(x)] + \bar{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbb{C}ov[g(\mathcal{X})]\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\bar{\Theta}(x,\mathcal{X})^T && (59) \\
& - 2\bar{\Theta}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbb{C}ov[g(\mathcal{X}), g(x)] && (60) \\
& && (61)
\end{aligned}
$$

For any Θ_{θ_0}, and f centered and decorrelated, by defining $g(x) = \Theta_{\theta_0}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)$, we get

$$
\begin{aligned}
\mathbb{V}[f(x)] = & \mathbb{V}[\Theta_{\theta_0}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] + \mathbb{V}[\bar{\mathcal{Q}}(x,\mathcal{X})\Theta_{\theta_0}(\mathcal{X},\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)] && (62) \\
& - 2\mathbb{C}ov[\bar{\mathcal{Q}}(x,\mathcal{X})\Theta_{\theta_0}(\mathcal{X},\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0), \Theta_{\theta_0}(x,\mathcal{X})\bar{\Theta}(\mathcal{X},\mathcal{X})^{-1}f(\mathcal{X},\theta_0)]. && (63)
\end{aligned}
$$

which is identical to $\mathbb{V}^i(x)$ from Proposition 2.1.

A.3.2 Noise correlation

In general, the noise in the NTK and in the functional initialization are related, as they come from the same weight initialization θ_0. Therefore, the analytical expression in Proposition 2.2 would have another covariance term. This covariance term disappears as we consider the models $f^{\text{gd-c}}$ and $f^{\text{gd-a}}$ defined in Section 3.2.2, which still manage to describe the predictive variance of the full model f as the interpolation of the two.

We can nevertheless construct a model such that the 2 noises are decorrelated for validating the taylor expansion from Prop 2.1, by sampling 2 initialization independently, θ_0^f and θ_0^k, and defining the model as such:

$$f^{\text{lin-d}}(x) = f(x, \theta_0^f) + \mathcal{Q}_{\theta_0^k}(x, \mathcal{X})(\mathcal{Y} - f(\mathcal{X}, \theta_0^f)) \tag{64}$$

We use such $f^{\text{lin-d}}$ to compute the predictive variance in Fig. 2 and 4.

A.4 Moments of the Neural Tangent Kernel

We consider the MLP defined in Section 2. We will consider the case where the output dimension h_L is 1 for ease of notation, but the derivations can be trivially extended into the multiple dimension case. The NTK is defined as:

$\forall x, x' \in \mathbb{R}^d$,

$$\Theta(x, x') := \nabla_\theta f(x) \nabla_\theta f(x')^T = \sum_{l=1}^{L} \sum_{i=1}^{h_l} \big[\nabla_{w_{l,i}} f(x) \nabla_{w_{l,i}} f(x')^T + \frac{\partial f}{\partial b_{l,i}}(x) \frac{\partial f}{\partial b_{l,i}}(x')\big] \tag{65}$$

where $w_{l,i}$ is the i-th column of W_l, and $b_{l,i}$ the i-th element of b_l.

Denoting by $z_{l,i}$ the i-th pre-activation in layer l for the input x, we have

$$\nabla_{w_{l,i}} f = \frac{\partial f}{\partial z_{l,i}} \nabla_{w_{l,i}} z_{l,i} = \frac{\partial f}{\partial b_{l,i}} \frac{\sigma_w}{\sqrt{h_{l-1}}} x_{l-1}^T$$

And thus

$$\begin{aligned}
\Theta(x, x') &= \sum_{l=1}^{L} \sum_{i=1}^{h_l} (\frac{\sigma_w^2}{h_{l-1}} x_{l-1}^T x'_{l-1} + 1) \frac{\partial f}{\partial b_{l,i}}(x) \frac{\partial f}{\partial b_{l,i}}(x') \\
&= 1 + \frac{\sigma_w^2}{h_{L-1}} x_{L-1}^T x'_{L-1} + \sum_{l=1}^{L-1} (\frac{\sigma_w^2}{h_{l-1}} x_{l-1}^T x'_{l-1} + 1) \sum_{i=1}^{h_l} \phi'(z_{l,i}) \phi'(z'_{l,i}) \frac{\partial f}{\partial x_{l,i}}(x) \frac{\partial f}{\partial x_{l,i}}(x')
\end{aligned} \tag{66}$$

where the x'_l, z'_l are the counterparts of x_l and z_l evaluated at x'.

For the single hidden layer case (e.g. $L = 2$), the above expression can be simplified into

$$\Theta(x, x') = 1 + \frac{\sigma_w^2}{h} \sum_{i=1}^{h} \big[\phi(z_i)\phi(z'_i) + (1 + \frac{\sigma_w^2}{d} x^T x') w_{2,0,i}^2 \phi'(z_i)\phi'(z'_i)\big] \tag{67}$$

where $h = h_1, d = h_0$, $w_{2,0,i}$ the i-th element of $w_{2,0}$, and $z_i = z_{1,i}$.

A.4.1 NTK first moment for 1 hidden layer MLP

For a single hidden MLP, we have

$$\mathbb{E}[\Theta(x, x')] = 1 + \mathbb{E}\big[\phi(z)\phi(z')\big] + (1 + \frac{\sigma_w^2}{d} x^T x') \mathbb{E}\big[\phi'(z)\phi'(z')\big] \tag{68}$$

which is identical to the infinite width deterministic NTK [24].

A.4.2 NTK second moment for 1 hidden layer MLP

Here, we assume the network to be a 1 hidden layer MLP. We then have

$$\Theta(x,x') = 1 + \frac{\sigma_w^2}{h}\sum_{i=1}^{h}\big[\phi(z_i)\phi(z_i') + (1+\frac{\sigma_w^2}{d}x^Tx')w_{2,0,i}^2\phi'(z_i)\phi'(z_i')\big] \tag{69}$$

and thus, using $\mathbb{C}ov(z_i,z_j)=0$, $\mathbb{C}ov(w_{2,0,i},w_{2,0,j})=0$ for any $i \neq j$ $\forall x,x',x'',x''' \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{C}ov[\Theta(x,x'),\Theta(x'',x''')] \\
=&\frac{\sigma_w^4}{h^2}\sum_{i=1}^{h}\mathbb{C}ov\big[\phi(z_i)\phi(z_i') + (1+\frac{\sigma_w^2}{d}x^Tx')w_{2,0,i}^2\phi'(z_i)\phi'(z_i'),\phi(z_i'')\phi(z_i''') \\
&+(1+\frac{\sigma_w^2}{d}x''^Tx''')w_{2,0,i}^2\phi'(z_i'')\phi'(z_i''')\big]
\end{aligned} \tag{70}$$

Since the z_i and $w_{2,0,i}$ are further identically distributed, by denoting by z, w the random variables respectively drawn from the same distributions,

$$\begin{aligned}
&\mathbb{C}ov[\Theta(x,x'),\Theta(x'',x''')] \\
=&\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z') + (1+\frac{\sigma_w^2}{d}x^Tx')w^2\phi'(z)\phi'(z'),\phi(z'')\phi(z''') \\
&+(1+\frac{\sigma_w^2}{d}x''^Tx''')w^2\phi'(z'')\phi'(z''')\big] \\
=&\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z'),\phi(z'')\phi(z''')\big] \\
&+(1+\frac{\sigma_w^2}{d}x^Tx')(1+\frac{\sigma_w^2}{d}x''^Tx''')\frac{\sigma_w^4}{h}\mathbb{C}ov\big[w^2\phi'(z)\phi'(z'),w^2\phi'(z'')\phi'(z''')\big] \\
&+(1+\frac{\sigma_w^2}{d}x''^Tx''')\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z'),w^2\phi'(z'')\phi'(z''')\big] \\
&+(1+\frac{\sigma_w^2}{d}x^Tx')\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z'')\phi(z'''),w^2\phi'(z)\phi'(z')\big] \\
=&\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z'),\phi(z'')\phi(z''')\big] \\
&+(1+\frac{\sigma_w^2}{d}x^Tx')(1+\frac{\sigma_w^2}{d}x''^Tx''')\frac{\sigma_w^4}{h}\big[3\mathbb{C}ov[\phi'(z)\phi'(z'),\phi'(z'')\phi'(z''')] \\
+&2\mathbb{E}[\phi'(z)\phi'(z')]\mathbb{E}[\phi'(z'')\phi'(z''')]\big] \\
&+(1+\frac{\sigma_w^2}{d}x''^Tx''')\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z'),\phi'(z'')\phi'(z''')\big] \\
&+(1+\frac{\sigma_w^2}{d}x^Tx')\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z'')\phi(z'''),\phi'(z)\phi'(z')\big]
\end{aligned} \tag{71}$$

In particular, we have

$$\begin{aligned}
\mathbb{V}[\Theta(x,x')] =&\frac{\sigma_w^4}{h}\mathbb{V}\big[\phi(z)\phi(z')\big] \\
&+(1+\frac{\sigma_w^2}{d}x^Tx')^2\frac{\sigma_w^4}{h}\big[3\mathbb{V}[\phi'(z)\phi'(z')]+2\mathbb{E}[\phi'(z)\phi'(z')]^2\big] \\
&+\frac{2\sigma_w^4}{h}(1+\frac{\sigma_w^2}{d}x^Tx')\mathbb{C}ov\big[\phi(z)\phi(z'),\phi'(z)\phi'(z')\big]
\end{aligned} \tag{72}$$

$$
\begin{aligned}
\mathbb{C}ov[\Theta(x,x'),\Theta(x,x)] =& \frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z'),\phi(z)^2\big] \\
&+ (1+\frac{\sigma_w^2}{d}x^Tx')(1+\frac{\sigma_w^2}{d}x^Tx)\frac{\sigma_w^4}{h}\big[3\mathbb{C}ov[\phi'(z)\phi'(z'),\phi'(z)^2] \\
&+ 2\mathbb{E}[\phi'(z)\phi'(z')]\mathbb{E}[\phi'(z)^2]\big] \\
&+ (1+\frac{\sigma_w^2}{d}x^Tx')\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)^2,\phi'(z)\phi'(z')\big] \\
&+ (1+\frac{\sigma_w^2}{d}x^Tx)\frac{\sigma_w^4}{h}\mathbb{C}ov\big[\phi(z)\phi(z'),\phi'(z)^2\big]
\end{aligned} \tag{73}
$$

A.4.3 Special case of ReLU activation

We now give the analytical expression of the first and second moments of the NTK for the 1 hidden layer MLP ReLU activation that are required to compute the predictive variance for a single training data setting. For simplicity, we assume the bias to be initialized to 0.

$\forall z \in \mathbb{R}$,

$$
\phi(z) = \mathbb{1}_{z>0}z \tag{74}
$$

$$
\phi'(z) = \mathbb{1}_{z>0} \tag{75}
$$

Following the previous notation, given x, the hidden activations are i.i.d. random variables $z = \frac{\sigma_w}{\sqrt{d}}w_1^Tx$ where w_1 is a univariate standard Gaussian random variable. σ_w is typically chosen to be $\sqrt{2}$ for ReLU activations.

We can rewrite a multivariate standard gaussian random variable as $w_1 = r.u$ where $r = ||w_1||_2$ is a real valued random variable distributed such that its squared value follows the Chi-squared distribution of degree $dim(w_1) = d$ and $u = \frac{w_1}{||w_1||_2}$ is a multivariate random variable uniformly distributed on the unit sphere. The 2 random variables are furthermore independent.

Let $x, x' \in \mathbb{R}^d$. We denote by $\theta = \arccos(\frac{x^Tx'}{||x||||x'||})$ the angle between the vectors. We define $S_{x,x'} = \{u \in \mathbb{R}^d s.t. ||u||_2 = 1, u^Tx > 0, u^Tx' > 0\}$.

We then have

$$
\begin{aligned}
\phi(z)\phi(z') &= \mathbb{1}_{z>0}z\mathbb{1}_{z'>0}z' \\
&= \frac{2}{d}\mathbb{1}_{z>0}\mathbb{1}_{z'>0}w_1^Txw_1^Tx' \\
&= \frac{2}{d}\mathbb{1}_{u\in S_{x,x'}}r^2u^Txu^Tx' \\
&= \frac{2}{d}\mathbb{1}_{u\in S_{x,x'}}r^2u_{||}^Txu_{||}^Tx' \\
&= \frac{2||x||||x'||}{d}r^2||u_{||}||^2\mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}\cos(\beta+\frac{\theta}{2})\cos(\beta-\frac{\theta}{2})
\end{aligned} \tag{76}
$$

Where $u_{||}$ is the component of u which is in the 2-dimensional subspace spanned by x, x' if $\theta \neq 0$, and any 2-dimensional subspace including x otherwise. $\beta = sign(u_{||}^Ty).\arccos(u_{||}^T\frac{v}{||v||})$, with $v = \frac{x}{||x||} + \frac{x'}{||x'||}$ and y a unit vector in the subspace orthogonal to v, is its angle in the subspace, uniformly distributed on $[-\pi,\pi]$. $\phi(z)\phi(z')$ is thus the product of 3 independent distribution, a random variable from a Chi-squared distribution of degree d, another one which depends on β, and finally on $||u_{||}||$.

Furthermore we have

$$
\begin{aligned}
\phi'(z)\phi'(z') &= \mathbb{1}_{z>0}\mathbb{1}_{z'>0} \\
&= \mathbb{1}_{u\in S_{x,x'}} \\
&= \mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}
\end{aligned} \tag{77}
$$

$\phi'(z)\phi'(z')$ is thus a Bernouilli distribution of probability $p = \frac{\pi-\theta}{2\pi}$.

Let us now compute the various quantities required for the predictive variance:

$$\begin{aligned}\mathbb{E}\big[\phi(z)\phi(z')\big] =& 2\frac{||x||||x'||}{d}\mathbb{E}[r^2]\mathbb{E}[||u_\parallel||^2]\mathbb{E}[\mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}\cos(\beta+\frac{\theta}{2})\cos(\beta-\frac{\theta}{2})] \\ =& ||x||||x'||C_d\frac{1}{2\pi}[(\pi-\theta)\cos(\theta)+\sin(\theta)]\end{aligned} \tag{78}$$

where we used $\cos(\beta+\frac{\theta}{2})\cos(\beta-\frac{\theta}{2}) = \frac{1}{2}[\cos(2\beta)+\cos(\theta)]$ and $\mathbb{E}(\chi^2_d) = d$.

$$\begin{aligned}\mathbb{V}\big[\phi(z)\phi(z')\big] =& 4\frac{||x||^2||x'||^2}{d^2}\mathbb{E}[r^4]\mathbb{E}[||u_\parallel||^4]\mathbb{E}[\mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}\cos^2(\beta+\frac{\theta}{2})\cos^2(\beta-\frac{\theta}{2})] \\ & - 4\frac{||x||^2||x'||^2}{d^2}\mathbb{E}[r^2]^2\mathbb{E}[||u_\parallel||^2]^2\mathbb{E}[\mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}\cos(\beta+\frac{\theta}{2})\cos(\beta-\frac{\theta}{2})]^2 \\ =& \frac{||x||^2||x'||^2}{d^2}(2d+d^2)C'_d\frac{1}{4\pi}\big[\frac{3}{2}\sin(2\theta)+(\pi-\theta)(\cos(2\theta)+2)\big] \\ & - ||x||^2||x'||^2C_d^2\frac{1}{4\pi^2}[(\pi-\theta)\cos(\theta)+\sin(\theta)]^2\end{aligned} \tag{79}$$

where we used $\mathbb{V}(\chi^2_d) = 2d$ and $C_d = \mathbb{E}[||u_\parallel||^2] = \frac{2}{d}$, $C'_d = \mathbb{E}[||u_\parallel||^4] = \frac{8}{2d+d^2}$.

Likewise,

$$\mathbb{E}\big[\phi'(z)\phi'(z')\big] = \frac{\pi-\theta}{2\pi} \tag{80}$$

$$\mathbb{V}\big[\phi'(z)\phi'(z')\big] = \frac{\pi-\theta}{2\pi}(1-\frac{\pi-\theta}{2\pi}) \tag{81}$$

as given by the Bernouilli distribution.

Finally,

$$\mathbb{C}ov\big[\phi(z)\phi(z'),\phi'(z)\phi'(z')\big] = \mathbb{E}\big[\phi(z)\phi(z')\big](1-\mathbb{E}[\phi'(z)\phi'(z')]) \tag{82}$$

and by using $\mathbb{1}_{u\in S_{x,x'}}\mathbb{1}_{u\in S_{x,x}} = \mathbb{1}_{u\in S_{x,x'}}$ as well as $\mathbb{E}[\phi'(z)^2] = \frac{1}{2}$,

$$\mathbb{C}ov\big[\phi(z)\phi(z'),\phi'(z)^2\big] = \mathbb{E}\big[\phi(z)\phi(z')\big](1-\mathbb{E}[\phi'(z)^2]) \tag{83}$$

$$\begin{aligned}\mathbb{C}ov\big[\phi(z)^2,\phi'(z)\phi'(z')\big] =& \mathbb{E}\big[\frac{2}{d}r^2(u_\parallel^Tx)^2\mathbb{1}_{u\in S_{x,x'}}\big] - \mathbb{E}\big[\phi(z)^2\big]\mathbb{E}\big[\mathbb{1}_{u\in S_{x,x'}}\big] \\ =& 2\|x\|^2C_d\mathbb{E}\big[\mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}\cos^2(\beta-\frac{\theta}{2})\big] - \frac{\|x\|^2C_d}{2}\frac{\pi-\theta}{2\pi} \\ =& \frac{\|x\|^2C_d}{4\pi}\big[2(\pi-\theta)+\sin(2\theta)-(\pi-\theta)\big] \\ =& \frac{\|x\|^2C_d}{4\pi}\big[(\pi-\theta)+\sin(2\theta)\big]\end{aligned} \tag{84}$$

$$\begin{aligned}\mathbb{C}ov\big[\phi(z)\phi(z'),\phi(z)^2\big] =& \mathbb{E}\big[\frac{4}{d^2}r^4(u_\parallel^Tx)^3u_\parallel^Tx'\mathbb{1}_{u\in S_{x,x'}}\big] - \mathbb{E}\big[\phi(z)^2\big]\mathbb{E}\big[\phi(z)\phi(z')\big] \\ =& 4\frac{2d+d^2}{d^2}\|x\|^3\|x'\|C'_d\mathbb{E}\big[\mathbb{1}_{\beta\in[-\frac{\pi-\theta}{2},\frac{\pi-\theta}{2}]}\cos^3(\beta-\frac{\theta}{2})\cos(\beta+\frac{\theta}{2})\big] \\ & - \|x\|^2C_d\|x\|\|x'\|C_d\frac{1}{4\pi}[(\pi-\theta)\cos(\theta)+\sin(\theta)] \\ =& \frac{2d+d^2}{16d^2\pi}\|x\|^3\|x'\|C'_d\big[\sin(3\theta)+9\sin(\theta)+12(\pi-\theta)\cos(\theta)\big] \\ & - \frac{\|x\|^3\|x'\|C_d^2}{4\pi}[(\pi-\theta)\cos(\theta)+\sin(\theta)]\end{aligned} \tag{85}$$

$$\begin{aligned}\mathbb{C}ov\big[\phi'(z)\phi'(z'), \phi'(z)^2\big] &= \mathbb{E}\big[\phi'(z)\phi'(z')\big](1-\mathbb{E}[\phi'(z)^2]) \\ &= \frac{\pi-\theta}{4\pi}\end{aligned} \tag{86}$$

Putting everything together in eq 68,72 and 73, and using

$$\bar{\mathcal{K}}(x,x') = \mathbb{E}\big[\phi(z)\phi(z')\big] = ||x||||x'||\frac{1}{d\pi}[(\pi-\theta)\cos(\theta)+\sin(\theta)] \tag{87}$$

gives us the analytical expression of the following variance terms:

$$\mathbb{V}[f^{lin-a}(x')] = \bar{\mathcal{K}}(x',x') - 2\frac{\mathbb{E}[\Theta(x',x)]}{\mathbb{E}[\Theta(x,x)]}\bar{\mathcal{K}}(x,x') + \frac{\mathbb{E}[\Theta(x',x)]^2}{\mathbb{E}[\Theta(x,x)]^2}\bar{\mathcal{K}}(x,x) \tag{88}$$

$$\begin{aligned}\mathbb{V}[f^{lin-c}(x')] =& \frac{1}{\mathbb{E}[\Theta(x,x)]^2}\mathbb{V}[\Theta(x',x)] + \frac{\mathbb{E}[\Theta(x',x)]^2}{\mathbb{E}[\Theta(x,x)]^4}\mathbb{V}[\Theta(x,x)] \\ &- 2\frac{\mathbb{E}[\Theta(x',x)]}{\mathbb{E}[\Theta(x,x)]^3}\mathbb{C}ov[\Theta(x,x),\Theta(x,x')]\end{aligned} \tag{89}$$

$$\begin{aligned}\mathbb{V}[f^{lin-i}(x')] =& \frac{1}{\mathbb{E}[\Theta(x,x)]^2}\mathbb{V}[\Theta(x',x)f(x)] + \frac{\mathbb{E}[\Theta(x',x)]^2}{\mathbb{E}[\Theta(x,x)]^4}\mathbb{V}[\Theta(x,x)f(x)] \\ &- 2\frac{\mathbb{E}[\Theta(x',x)]}{\mathbb{E}[\Theta(x,x)]^3}\mathbb{C}ov[\Theta(x,x)f(x),\Theta(x,x')f(x)] \\ =& \bar{\mathcal{K}}(x,x)\mathbb{V}[f^{lin-c}(x')]\end{aligned} \tag{90}$$

In particular, when $\|x\|, \|x'\| << \sqrt{d}$, the first-order approximation of $\mathbb{V}[f^{lin-c}(x')]$ becomes a function which only depends on θ, which could be seen on Fig. 1. We analytically validate the expression in Fig. 7.

B Appendix: empirical results

In this Appendix Section, we provide more data on similar experiments described in Section 2 and 3 of the manuscript. Generally, we conducted our experiments on 4 Linux servers with 8 Nvidia RTX 3090 GPUs with 24 GB memory each. The presented experiments are compute-intensive which led to experiments validating our theoretical propositions on rather small networks and datasets. During the development, we conducted many scans over ensemble width and depth as well as datasets over the course of several months. Despite heavily relying on PyTorch, we conducted NTK kernel experiments with the following Github codebase. We thank the authors for providing this excellent resource (https://github.com/google/neural-tangents).

Further details about our general setup and training specifications are not described in the text. Missing details may be described in the accompanied code.

- We choose a learning rate $\eta = 0.1$ and trained all of our models with gradient descent and momentum (0.9) for all (linearized) training experiments. Although the learning rate is relatively high, we saw that the models trained with gradient descent align very well with the kernel models.
- For the CNN, we always use filter size of 3 and padding. Every 2nd layer, we use a stride of 2. Before the last layer, we flatten the features and linearly project to the output. We always use the NTK initialization as introduced above.
- For the SGD results in Table 1, we used a batchsize of 1000.
- Whenever we used kernel models and a small dataset (N=100), we restricted the problem to be a binary classification problem.
- For the WRN 28-10 experiments, we used learning rate $\eta = 0.03$ and batchsize 128 the standard network specifications as in https://github.com/hysts/pytorch_wrn, with momentum. We train the model for 10 epochs for the cross entropy loss, and 30 for the MSE loss.
- For all models trained on the MSE loss, we use as target the centered one-hot encoding of the class variable, as in [31].
- For the AUROC computation, we used the standard method from the `SciPy` package.

B.1 Additional empirical results

The following results are presented in the Appendix:

- Figure 5 and Table 3: Confirmation of the assumption used in Proposition 2.3 for MLP and CNN respectively trained on a subset of MNIST and CIFAR10.
- Figure 6: $\mathcal{R}(f)$ for CNNs trained on a subset of CIFAR10 in support of Proposition 2.3.
- Table 4: AUROC for all OOD datasets i.e. SVHN, LSUN, TIN, iSUN, CIFAR100.
- Table 5: Predictive variance (on test set), test set accuracy and AUROC for kernel as well as models trained with gradient descent on full MNIST (N=50000). The same trend as for N=1000 is observed i.e. the gradient descent ensembles follow closely the linearly trained ensembles behavior.
- Table 6: Test set accuracy and AUROC for (stochastic) linearly trained models as well as models trained with (stochastic) gradient descent on a subset of MNIST (N=1000). We observe tiny differences between the stochastic and its non-stochastic counterpart.
- Table B.1: Test set accuracy and AUROC for WRN 28-10 ensembles of size 8 trained on CIFAR100. We trained the models with the cross entropy (CE) and MSE loss, for respectively 10 and 30 epochs. For the MSE loss, the network output was regressed against the one-hot encoding of the target class, centered to be of 0 mean and rescaled by a factor 10.
- Table B.1: Test set accuracy and AUROC for an AlexNet ensembles of size 8 trained on FashionMNIST, with the cross entropy (CE) loss, for 50 epochs, with momentum.

For computing the AUROC values that play a central part in our empirical evaluation we simply collect predictions from in-distribution i.e. the test dataset of the corresponding training dataset as

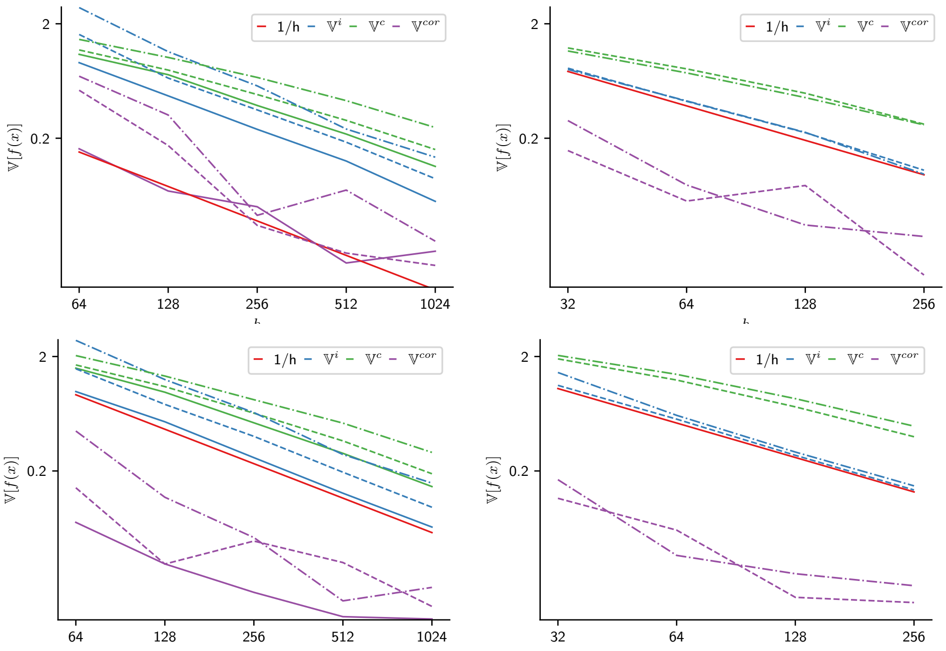

Figure 4: Verification of the scaling of $1/h$ of the variance terms influenced by the kernel noise as well as $1/h^2$ of the residual. Although the theoretical result holds only for depth $L = 2$ (line plots), the same scaling is observed for deeper networks as suggested by our informal result ($L = 3$ in dashed lines, $L = 5$ in dashed-dotted lines). *Upper Row:* Predictive variance $\mathbb{V}^i, \mathbb{V}^c$, as well as $\mathbb{V}^{cor}$ of an ensemble of MLPs (left) and CNN (right) of various depths and widths trained on a subset of MNIST (N=100). *Lower Row:* Predictive variance $\mathbb{V}^i, \mathbb{V}^c$ as well as $\mathbb{V}^{cor}$ of an ensemble of MLPs (left) and CNN (right) of various depths and widths trained on a subset of CIFAR10 (N=100).

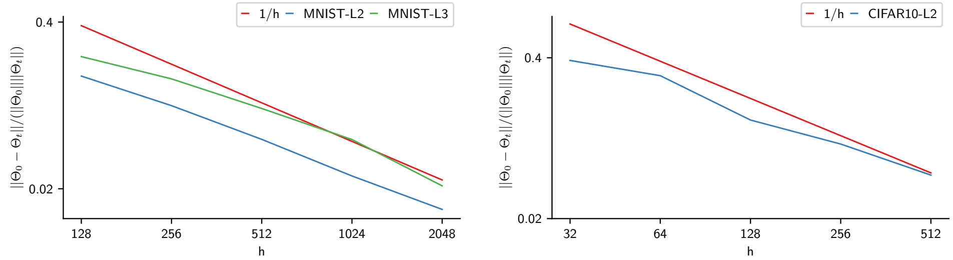

Figure 5: Verification of scaling of $1/h$ of for the relative neural tangent kernel change. *Left:* MLPs on a subset of MNIST (N=100) for depth $L \in \{2, 3\}$. *Right:* CNNs on a subset of CIFAR10 (N=100) for depth $L = 2$. All plots are in log-log scale.

well as predictions from the out-of-distribution datasets which vary across setups, see above. To compute per in- and out-of-distribution pair, we compute the auroc values with the help of the publicly available `sklearn` package and its `metrics.roc_auc_score` function. We report the average over the pairs.

Table 3: Further verification of scaling of $1/h$ of for the relative neural tangent kernel change. $\|\Theta_0 - \Theta_t\|/\|\Theta_0\|$ for trained MLPs on a subset of MNIST (N=500) for depth $L \in \{2, 3\}$.

Depth	Width 512	1024	2048	4096
2	0.2214	0.1406	0.0676	0.0368
3	0.3500	0.2218	0.1226	0.0701

Table 4: Predictive variance, test set accuracy and AUROC for deep ensembles of size 10 of CNNs ($h = 256$, $L = 3$) trained an a subset (N=1000) of CIFAR10. We indicate small standard deviations σ obtained over 3 ensembles of size E with $\pm.00$. In all experiments, the various disentangled models show significant differences in behavior. All linearly trained models follow the gradient descent models behavior tightly. When optimizing with SGD, isolating initial noise sources still affect the ensemble behavior significantly and can lead to improved OOD detection as well as test set accuracy in this more realistic settings.

Model	CNN, CIFAR10, N=1000, E=10, η=0.1						
	$\mathbb{V}$	Test (%)	C100	SVHN	LSUN	TIN	iSUN
f^{lin}	$0.400^{\pm 0.005}$	$36.43^{\pm.90}$	$0.537^{\pm.005}$	$.532^{\pm.006}$	$.809^{\pm.004}$	$0.796^{\pm.003}$	$.783^{\pm.004}$
$f^{\text{lin-c}}$	$0.106^{\pm 0.001}$	$37.20^{\pm.44}$	$0.535^{\pm.002}$	$.567^{\pm.006}$	$.693^{\pm.001}$	$0.689^{\pm.003}$	$.674^{\pm.004}$
$f^{\text{lin-a}}$	$1.277^{\pm 0.051}$	$30.90^{\pm.53}$	$0.526^{\pm.002}$	$.510^{\pm.006}$	$.764^{\pm.003}$	$0.749^{\pm.004}$	$.738^{\pm.000}$
$f^{\text{lin-i}}$	$0.443^{\pm 0.008}$	$32.85^{\pm.21}$	$0.531^{\pm.001}$	$.591^{\pm.003}$	$.683^{\pm.001}$	$0.681^{\pm.004}$	$.660^{\pm.000}$
f^{gd}	$0.442^{\pm 0.004}$	$39.70^{\pm.52}$	$0.534^{\pm.003}$	$.516^{\pm.002}$	$.789^{\pm.003}$	$0.774^{\pm.001}$	$.763^{\pm.004}$
$f^{\text{gd-c}}$	$0.112^{\pm 0.001}$	$37.47^{\pm.49}$	$0.535^{\pm.002}$	$.562^{\pm.004}$	$.691^{\pm.004}$	$0.683^{\pm.003}$	$.670^{\pm.002}$
$f^{\text{gd-a}}$	$1.316^{\pm 0.045}$	$30.53^{\pm 1.15}$	$0.527^{\pm.002}$	$.509^{\pm.004}$	$.758^{\pm.005}$	$0.746^{\pm.004}$	$.734^{\pm.003}$
$f^{\text{gd-i}}$	$0.505^{\pm 0.004}$	$31.20^{\pm.14}$	$0.524^{\pm.000}$	$.583^{\pm.000}$	$.656^{\pm.003}$	$0.654^{\pm.007}$	$.638^{\pm.003}$
Train	CIFAR10, N=50000, E=5, batchsize=1000, η=0.1						
	$\mathbb{V}$	Test (%)	C100	SVHN	LSUN	TIN	iSUN
f^{sgd}	$.03^{\pm.00}$	$62.68^{\pm.36}$	$.557^{\pm.00}$	$.557^{\pm.01}$	$.884^{\pm.00}$	$.878^{\pm.00}$	$.864^{\pm.00}$
$f^{\text{sgd-c}}$	$.01^{\pm.00}$	$57.03^{\pm.14}$	$.548^{\pm.00}$	$.554^{\pm.00}$	$.791^{\pm.00}$	$.791^{\pm.00}$	$.781^{\pm.00}$
$f^{\text{sgd-a}}$	$.19^{\pm.01}$	$58.83^{\pm.22}$	$.536^{\pm.00}$	$.455^{\pm.00}$	$.864^{\pm.00}$	$.858^{\pm.00}$	$.845^{\pm.00}$

Table 5: Test set accuracy and AUROC for deep ensembles of size 5 of MLPs ($h = 1024$, $L = 3$) trained on full MNIST. We indicate small standard deviations σ obtained over 3 ensembles with $\pm.00$. In all experiments, the various disentangled models show significant differences in behavior. All linearly trained models follow the gradient descent models behavior tightly even in this regime of full MNIST.

Model	MLP, MNIST, N=50000, η=0.1				
	$\mathbb{V}$	Test (%)	FM	EM	KM
f^{sgd}	$.08^{\pm.00}$	$95.7^{\pm.1}$	$.974^{\pm.01}$	$.930^{\pm.00}$	$.991^{\pm.00}$
$f^{\text{sgd-c}}$	$.01^{\pm.00}$	$94.4^{\pm.0}$	$.924^{\pm.02}$	$.873^{\pm.01}$	$.962^{\pm.00}$
$f^{\text{sgd-a}}$	$.22^{\pm.03}$	$97.5^{\pm.1}$	$.988^{\pm.00}$	$.943^{\pm.00}$	$.995^{\pm.00}$
f^{lin}	$.05^{\pm.00}$	$96.5^{\pm.1}$	$.965^{\pm.01}$	$.986^{\pm.00}$	$.995^{\pm.00}$
$f^{\text{lin-c}}$	$.01^{\pm.00}$	$94.4^{\pm.0}$	$.923^{\pm.01}$	$.872^{\pm.02}$	$.965^{\pm.00}$
$f^{\text{lin-a}}$	$.23^{\pm.03}$	$97.8^{\pm.0}$	$.987^{\pm.00}$	$.940^{\pm.00}$	$.993^{\pm.00}$

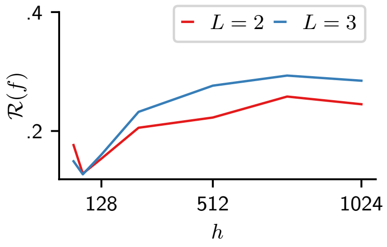

Figure 6: $\mathcal{R}(f)$ of CNNs with multiple widths and depths ($L \in \{2, 3\}$) trained on a subset (100) on CIFAR10. As predicted, we observe that $\mathcal{R}(f)$ remains bounded as the width increases.

Table 6: Comparison of gradient and stochastic gradient descent for deep ensembles of MLPs ($h = 1024$, $L = 3$) trained on a subset (N=1000) of MNIST. For the linearly trained models, gradients for the different batches are computed at initialization and applied stochastically. The two different optimization methods lead to practically indistinguishable behavior measured through test set accuracy and AUROC. The batchsize is set to 100 for the sgd models.

Model	MLP, MNIST, N=1000, η=0.1			
	Test	FM	EM	KM
$f^{\text{sgd-lin}}$	91.60	.968	.922	.982
$f^{\text{sgd-lin-c}}$	89.70	.930	.879	.965
$f^{\text{sgd-lin-a}}$	91.20	.980	.923	.987
f^{sgd}	91.10	.976	.924	.986
$f^{\text{sgd-c}}$	89.70	.932	.882	.966
$f^{\text{sgd-a}}$	90.50	.981	.923	.988
f^{lin}	91.60	.968	.922	.982
$f^{\text{lin-c}}$	89.70	.930	.879	.965
$f^{\text{lin-a}}$	91.20	.980	.923	.987
f^{gd}	91.10	.976	.923	.986
$f^{\text{gd-c}}$	89.70	.932	.882	.966
$f^{\text{gd-a}}$	90.50	.981	.923	.988

Table 7: Test set accuracy and AUROC for WRN 28-10 ensembles of size 8 trained on CIFAR100, with the cross entropy (CE) and MSE loss. Standard deviations σ computed over 5 seeds are indicated with ±. In bold are values that statistically significantly outperform f^{sgd} with $p < 0.2$.

Model	WRN 28-10, CIFAR100, batchsize 128, η=0.03					
	Test (%)	C10	SVHN	LSUN	TIN	iSUN
$f^{\text{sgd}}(CE)$	$67.57^{\pm 0.37}$	$0.703^{\pm 0.003}$	$0.776^{\pm 0.005}$	$0.735^{\pm 0.003}$	$0.742^{\pm 0.004}$	$0.741^{\pm 0.003}$
$f^{\text{sgd-c}}(CE)$	$67.59^{\pm 0.35}$	$\mathbf{0.708}^{\pm 0.003}$	$0.778^{\pm 0.004}$	$\mathbf{0.738}^{\pm 0.004}$	$0.744^{\pm 0.005}$	$0.744^{\pm 0.006}$
$f^{\text{sgd-a}}(CE)$	$67.26^{\pm 0.11}$	$0.705^{\pm 0.003}$	$0.773^{\pm 0.003}$	$0.735^{\pm 0.003}$	$0.741^{\pm 0.001}$	$0.742^{\pm 0.002}$
$f^{\text{sgd}}(MSE)$	$63.00^{\pm 0.15}$	$0.704^{\pm 0.003}$	$0.741^{\pm 0.005}$	$0.715^{\pm 0.004}$	$0.739^{\pm 0.005}$	$0.722^{\pm 0.005}$
$f^{\text{sgd-c}}(MSE)$	$62.90^{\pm 0.06}$	$0.705^{\pm 0.002}$	$\mathbf{0.746}^{\pm 0.004}$	$\mathbf{0.720}^{\pm 0.005}$	$\mathbf{0.743}^{\pm 0.004}$	$0.725^{\pm 0.006}$
$f^{\text{sgd-a}}(MSE)$	$62.19^{\pm 0.24}$	$\mathbf{0.710}^{\pm 0.004}$	$0.740^{\pm 0.009}$	$\mathbf{0.729}^{\pm 0.006}$	$\mathbf{0.749}^{\pm 0.004}$	$\mathbf{0.730}^{\pm 0.004}$

Table 8: Test set accuracy and AUROC for an AlexNet ensembles of size 5 trained on FashionMNIST, with the cross entropy (CE) loss. Standard deviations σ computed over 3 seeds are indicated with ±. In bold are values that statistically significantly outperform f^{sgd} with $p < 0.2$.

Model	AlexNet, FMNIST, batchsize 512, η=0.01			
	Test (%)	MNIST	EMNIST	KMNIST
$f^{\text{sgd}}(CE)$	$93.22^{\pm 0.39}$	$0.868^{\pm 0.014}$	$0.856^{\pm 0.004}$	$0.935^{\pm 0.006}$
$f^{\text{sgd-c}}(CE)$	$93.21^{\pm 0.13}$	$\mathbf{0.883}^{\pm 0.007}$	$\mathbf{0.867}^{\pm .011}$	$0.933^{\pm 0.005}$
$f^{\text{sgd-a}}(CE)$	$93.12^{\pm 0.09}$	$\mathbf{0.880}^{\pm 0.011}$	$0.838^{\pm 0.006}$	$0.926^{\pm 0.003}$

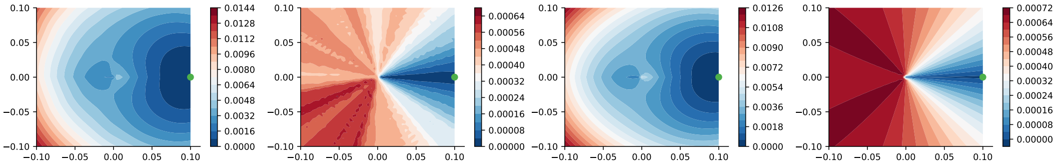

Figure 7: *Left 2 plots*: Empirically measured $\mathbb{V}^a$ (*left*) and $\mathbb{V}^c$ (*center left*) using an ensemble of size 100 linearly trained 1 hidden layer MLPs with 512 hidden units trained on a single datapoint (green point) with target 1 in the 2d space. *Right 2 plots*: analytically computed $\mathbb{V}^a$ (*center right*) and $\mathbb{V}^c$ (*right*) using the derivation in A.4.3.