Appendices

A. Proof of Lemma 3.1

We have $Q(\tilde{\mathbf{w}}(t))$ such that for all t, $\nabla Q(\tilde{\mathbf{w}}(t)) = \mathbf{X}\boldsymbol{\nu}(t)$. Thus we get that for every finite time t, $\tilde{\mathbf{w}}(t)$ is the solution of the minimization problem $\tilde{\mathbf{w}}(t) = \arg\min_{\mathbf{w}} Q(\mathbf{w}) \quad \text{s.t. } \mathbf{X}^\top \mathbf{w} = \mathbf{X}^\top \tilde{\mathbf{w}}(t)$. Since the square loss $\|\mathbf{X}^\top \tilde{\mathbf{w}}(t) - \mathbf{y}\|^2$ is minimized by gradient flow and therefore bounded for all time t, we get that the predictions are also bounded, i.e. $\forall t : \|\mathbf{X}^\top \tilde{\mathbf{w}}(t)\| < C$ for some finite C. Therefore, the solution $\mathbf{w}$ to the minimization problem above is of bounded constraints and is therefore also of bounded norm, i.e. $\|\tilde{\mathbf{w}}(t)\| < C'$ for some finite C' for all t. Taking the limit $t \to \infty$ we therefore converge to a finite weight vector $\tilde{\mathbf{w}}(\infty)$.

Next, from the relation $\nabla Q(\tilde{\mathbf{w}}(t)) = \mathbf{X}\boldsymbol{\nu}(t)$ we get that if the RHS is infinite at $t \to \infty$ then $\nabla Q(\tilde{\mathbf{w}}(\infty))$ is infinite. However, since we converge to a finite weight vector $\tilde{\mathbf{w}}(\infty)$, we get a contradiction since $\nabla Q(\tilde{\mathbf{w}})$ is bounded for any finite input. Therefore, $\lim_{t\to\infty} \mathbf{X}\boldsymbol{\nu}(t)$ is finite.

Next, we decompose $\mathbf{X}$ to its singular value decomposition

$$\mathbf{X} = \sum_{j:s_j>0} s_j \mathbf{v}_j \boldsymbol{u}_j^\top$$

where s_i are the singular values, and $\{\mathbf{v}_i\}$ and $\{\mathbf{u}_i\}$ are two sets of orthogonal vectors. We define

$$\boldsymbol{\nu} \triangleq \lim_{t\to\infty} \sum_{i:s_i>0} \boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{\nu}(t)$$

and note that

$$\begin{aligned}
\mathbf{X}\boldsymbol{\nu} &= \lim_{t\to\infty} \sum_{j:s_j>0} s_j \mathbf{v}_j \boldsymbol{u}_j^\top \sum_{i:s_i>0} \boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{\nu}(t) \\
&= \lim_{t\to\infty} \sum_{j:s_j>0} \sum_{i:s_i>0} s_j \mathbf{v}_i \boldsymbol{u}_j^\top \boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{\nu}(t) \\
&\overset{(1)}{=} \lim_{t\to\infty} \sum_{j:s_j>0} s_i \mathbf{v}_i \boldsymbol{u}_i^\top \boldsymbol{\nu}(t) \\
&= \lim_{t\to\infty} \mathbf{X}\boldsymbol{\nu}(t) = \nabla Q(\tilde{\mathbf{w}}(\infty))
\end{aligned}$$

where in (1) we used the fact that $\boldsymbol{u}_j^\top \boldsymbol{u}_i = \delta_{i,j}$. Therefore, $\boldsymbol{\nu}$ respects the KKT stationary condition

$$\mathbf{X}\boldsymbol{\nu} = \nabla Q(\tilde{\mathbf{w}}(\infty)) \,.$$

Lastly, we show that $\boldsymbol{\nu}$ is finite. Recall that

$$\nabla Q(\tilde{\mathbf{w}}(\infty)) = \lim_{t\to\infty} \sum_{j:s_j>0} s_i \boldsymbol{v}_i \boldsymbol{u}_i^\top \boldsymbol{\nu}(t)$$

For any i such that $s_i > 0$, if we multiply this equation by $\boldsymbol{v}_i^\top$ from the left, recalling $\mathbf{v}_i^\top \mathbf{v}_i = \delta_{i,j}$, we obtain

$$\infty > s_i^{-1} \boldsymbol{v}_i^\top \nabla Q(\tilde{\mathbf{w}}(\infty)) = \lim_{t\to\infty} \boldsymbol{u}_i^\top \boldsymbol{\nu}(t)$$

Therefore,

$$\|\boldsymbol{\nu}\| = \left\| \lim_{t\to\infty} \sum_{i:s_i>0} \boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{\nu}(t) \right\| < \infty \,.$$

B. Proof of Theorem 4.1

Proof. We examine a two-layer "diagonal linear network" with untied weights

$$f(\mathbf{x};\mathbf{u}_+,\mathbf{u}_-,\mathbf{v}_+,\mathbf{v}_-)=(\mathbf{u}_+\circ\mathbf{v}_+-\mathbf{u}_-\circ\mathbf{v}_-)^\top\mathbf{x}=\tilde{\mathbf{w}}^\top\mathbf{x}\,,$$

where

$$\tilde{\mathbf{w}}=\mathbf{u}_+\circ\mathbf{v}_+-\mathbf{u}_-\circ\mathbf{v}_-\,. \tag{17}$$

The gradient flow dynamics of the parameters is given by:

$$\frac{du_{+,i}}{dt}=-\frac{\partial\mathcal{L}}{\partial u_{+,i}}=v_{+,i}(t)\left(\sum_{n=1}^{N}x_i^{(n)}r^{(n)}(t)\right)$$
$$\frac{du_{-,i}}{dt}=-\frac{\partial\mathcal{L}}{\partial u_{-,i}}=-v_{-,i}(t)\left(\sum_{n=1}^{N}x_i^{(n)}r^{(n)}(t)\right)$$
$$\frac{dv_{+,i}}{dt}=-\frac{\partial\mathcal{L}}{\partial v_{+,i}}=u_{+,i}(t)\left(\sum_{n=1}^{N}x_i^{(n)}r^{(n)}(t)\right)$$
$$\frac{dv_{-,i}}{dt}=-\frac{\partial\mathcal{L}}{\partial v_{-,i}}=-u_{-,i}(t)\left(\sum_{n=1}^{N}x_i^{(n)}r^{(n)}(t)\right)$$

where we denote the residual

$$r^{(n)}(t)\triangleq y^{(n)}-\tilde{\mathbf{w}}^\top(t)\mathbf{x}^{(n)}\,.$$

From Eq. 17 we can write:

$$\begin{aligned}\frac{d\tilde{w}_i}{dt}&=\frac{du_{+,i}}{dt}v_{+,i}+u_{+,i}\frac{dv_{+,i}}{dt}-\frac{du_{-,i}}{dt}v_{-,i}-u_{-,i}\frac{dv_{-,i}}{dt}\\&=v_{+,i}^2\sum_{n=1}^{N}x_i^{(n)}r^{(n)}+u_{+,i}^2\sum_{n=1}^{N}x_i^{(n)}r^{(n)}+v_{-,i}^2\sum_{n=1}^{N}x_i^{(n)}r^{(n)}+u_{-,i}^2\sum_{n=1}^{N}x_i^{(n)}r^{(n)}\\&=\left(u_{+,i}^2+v_{+,i}^2+u_{-,i}^2+v_{-,i}^2\right)\sum_{n=1}^{N}x_i^{(n)}r^{(n)}\,.\end{aligned}$$

Thus,

$$\frac{1}{u_{+,i}^2+v_{+,i}^2+u_{-,i}^2+v_{-,i}^2}\frac{d\tilde{w}_i}{dt}=\sum_{n=1}^{N}x_i^{(n)}r^{(n)}\,.$$

We note that the quantity $u_{+,i}u_{-,i}+v_{+,i}v_{-,i}$ is conserved during training, since

$$\begin{aligned}\frac{d}{dt}\left(u_{+,i}u_{-,i}+v_{+,i}v_{-,i}\right)&=\frac{du_{+,i}}{dt}u_{-,i}+u_{+,i}\frac{du_{-,i}}{dt}+\frac{dv_{+,i}}{dt}v_{-,i}+v_{+,i}\frac{dv_{-,i}}{dt}\\&=u_{-,i}v_{+,i}\sum_{n=1}^{N}x_i^{(n)}r^{(n)}-u_{+,i}v_{-,i}\sum_{n=1}^{N}x_i^{(n)}r^{(n)}+u_{+,i}v_{-,i}\sum_{n=1}^{N}x_i^{(n)}r^{(n)}-v_{+,i}u_{-,i}\sum_{n=1}^{N}x_i^{(n)}r^{(n)}\\&=0\,.\end{aligned}$$

So

$$u_{+,i}u_{-,i}+v_{+,i}v_{-,i}=u_{+,i}\left(0\right)u_{-,i}\left(0\right)+v_{+,i}\left(0\right)v_{-,i}\left(0\right)\triangleq c_i\,. \tag{18}$$

Combining Eq. 17 and Eq. 18 we can write:

$$\begin{cases}\tilde{w}_i=u_{+,i}v_{+,i}-u_{-,i}v_{-,i}\\u_{+,i}u_{-,i}+v_{+,i}v_{-,i}=c_i\end{cases}\Rightarrow\begin{cases}\tilde{w}_i^2=u_{+,i}^2v_{+,i}^2+u_{-,i}^2v_{-,i}^2-2u_{+,i}v_{+,i}u_{-,i}v_{-,i}\\u_{+,i}^2u_{-,i}^2+v_{+,i}^2v_{-,i}^2+2u_{+,i}u_{-,i}v_{+,i}v_{-,i}=c_i^2\end{cases}$$

$$\Rightarrow u_{+,i}^2 u_{-,i}^2 + v_{+,i}^2 v_{-,i}^2 + u_{+,i}^2 v_{+,i}^2 + u_{-,i}^2 v_{-,i}^2 - \tilde{w}_i^2 = c_i^2 \ . \tag{19}$$

We also know that:

$$v_{+,i}^2 - u_{+,i}^2 = v_{+,i}^2(0) - u_{+,i}^2(0) \triangleq \delta_{+,i}$$
$$v_{-,i}^2 - u_{-,i}^2 = v_{-,i}^2(0) - u_{-,i}^2(0) \triangleq \delta_{-,i}$$

which can be easily shown since $\frac{d}{dt}\left(v_{+,i}^2 - u_{+,i}^2\right) = 0$ and $\frac{d}{dt}\left(v_{-,i}^2 - u_{-,i}^2\right) = 0$. So using Eq. 19 we can write:

$$u_{+,i}^2 u_{-,i}^2 + \left(\delta_{+,i} + u_{+,i}^2\right)\left(\delta_{-,i} + u_{-,i}^2\right) + u_{+,i}^2\left(\delta_{+,i} + u_{+,i}^2\right) + u_{-,i}^2\left(\delta_{-,i} + u_{-,i}^2\right) - \tilde{w}_i^2 = c_i^2$$
$$\Rightarrow \left(u_{+,i}^2 + u_{-,i}^2\right)^2 + \left(\delta_{+,i} + \delta_{-,i}\right)\left(u_{+,i}^2 + u_{-,i}^2\right) + \delta_{+,i}\delta_{-,i} - \tilde{w}_i^2 - c_i^2 = 0$$

$$\begin{aligned}\Rightarrow u_{+,i}^2 + u_{-,i}^2 &= \frac{-\left(\delta_{+,i} + \delta_{-,i}\right) + \sqrt{\left(\delta_{+,i} + \delta_{-,i}\right)^2 - 4\left(\delta_{+,i}\delta_{-,i} - \tilde{w}_i^2 - c_i^2\right)}}{2} \\ &= \frac{-\left(\delta_{+,i} + \delta_{-,i}\right) + \sqrt{\left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2 + 4\tilde{w}_i^2}}{2} \ .\end{aligned} \tag{20}$$

Coming back to $u_{+,i}^2 + v_{+,i}^2 + u_{-,i}^2 + v_{-,i}^2$ we have using Eq. 20 that:

$$\begin{aligned}u_{+,i}^2 + v_{+,i}^2 + u_{-,i}^2 + v_{-,i}^2 &= 2\left(u_{+,i}^2 + u_{-,i}^2\right) + \delta_{+,i} + \delta_{-,i} \\ &= \sqrt{\left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2 + 4\tilde{w}_i^2} \ .\end{aligned}$$

Therefore,

$$\frac{1}{\sqrt{\left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2 + 4\tilde{w}_i^2}} \frac{d\tilde{w}_i}{dt} = \sum_{n=1}^{N} x_i^{(n)} r^{(n)} \ .$$

We follow the IMD approach for deriving the implicit bias (presented in detail in Section 3 of the main paper) and try and find a function $q(\tilde{w}_i)$ such that:

$$\nabla^2 q\left(\tilde{w}_i(t)\right) = \frac{1}{\sqrt{\left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2 + 4\tilde{w}_i^2}} \ , \tag{21}$$

which will then give us that

$$\nabla^2 q\left(\tilde{w}_i(t)\right) \frac{d}{dt}\tilde{w}_i(t) = \sum_{n=1}^{N} x_i^{(n)} r^{(n)}$$

or

$$\frac{d}{dt}\left(\nabla q\left(\tilde{w}_i(t)\right)\right) = \sum_{n=1}^{N} x_i^{(n)} r^{(n)} \ .$$

Integrating the above, we get

$$\nabla q\left(\tilde{w}_i(t)\right) - \nabla q\left(\tilde{w}_i(0)\right) = \sum_{n=1}^{N} x_i^{(n)} \int_0^t r^{(n)}(t')dt' \ .$$

Denoting $\nu^{(n)} = \int_0^\infty r^{(n)}(t')dt'$, and assuming q also satisfies $\nabla q\left(\tilde{w}_i(0)\right) = 0$, will in turn give us the KKT stationarity condition

$$\nabla q\left(\tilde{w}_i(\infty)\right) = \sum_{n=1}^{N} x_i^{(n)} \nu^{(n)} \ .$$

Namely, if we find a q that satisfies the conditions above we will have that gradient flow (for each weight $\tilde{w}_i$) satisfies the KKT conditions for minimizing this q.

We next turn to solving for this q, beginning with Eq. 21:

$$q''\left(\tilde{w}_i\right) = \frac{1}{\sqrt{\left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2 + 4\tilde{w}_i^2}} = \frac{1}{\sqrt{k_i + 4\tilde{w}_i^2}} \ ,$$

where $k_i \triangleq \left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2$.

Integrating the above, and using the constraint $q'\left(0\right) = 0$ we get:

$$q'\left(\tilde{w}_i\right) = \frac{\log\left(\sqrt{4\tilde{w}_i^2 + k} + 2\tilde{w}_i\right) - \log\left(\sqrt{k}\right)}{2} \ .$$

Simplifying the above we obtain:

$$q'\left(\tilde{w}_i\right) = \frac{1}{2}\log\left(\frac{\sqrt{4\tilde{w}_i^2 + k_i} + 2\tilde{w}_i}{\sqrt{k_i}}\right) = \frac{1}{2}\log\left(\sqrt{1 + \frac{4\tilde{w}_i^2}{k_i}} + \frac{2\tilde{w}_i}{\sqrt{k_i}}\right) = \frac{1}{2}\text{arcsinh}\left(\frac{2\tilde{w}_i}{\sqrt{k_i}}\right) \ .$$

Finally, we integrate again to obtain the desired q:

$$q_{k_i}\left(\tilde{w}_i\right) = \frac{1}{2}\int_0^{\tilde{w}_i} \text{arcsinh}\left(\frac{2z}{\sqrt{k_i}}\right) dz = \frac{\sqrt{k_i}}{4}\left[1 - \sqrt{1 + \frac{4\tilde{w}_i^2}{k_i}} + \frac{2\tilde{w}_i}{\sqrt{k_i}}\text{arcsinh}\left(\frac{2\tilde{w}_i}{\sqrt{k_i}}\right)\right] \ ,$$

where

$$k_i = \left(\delta_{+,i} - \delta_{-,i}\right)^2 + 4c_i^2 = \left(v_{+,i}^2\left(0\right) - u_{+,i}^2\left(0\right) - v_{-,i}^2\left(0\right) + u_{-,i}^2\left(0\right)\right)^2 + 4\left(u_{+,i}\left(0\right)u_{-,i}\left(0\right) + v_{+,i}\left(0\right)v_{-,i}\left(0\right)\right)^2 \ .$$

For the case $u_{+,i}\left(0\right) = u_{-,i}\left(0\right), v_{+,i}\left(0\right) = v_{-,i}\left(0\right)$ (unbiased initialization of $\tilde{w}_i\left(0\right) = 0$) we get

$$k_i = 4\left(u_{+,i}^2\left(0\right) + v_{+,i}^2\left(0\right)\right)^2$$

$$\Rightarrow \sqrt{k_i} = 2\left(u_{+,i}^2\left(0\right) + v_{+,i}^2\left(0\right)\right) = \frac{4\alpha_i\left(1 + s_i^2\right)}{1 - s_i^2} \ .$$

Next, if we denote $Q_{\mathbf{k}}(\tilde{\mathbf{w}}) = \sum_{i=1}^d q_{k_i}\left(\tilde{w}_i\right)$, we can write

$$\nabla Q_{\mathbf{k}}(\tilde{\mathbf{w}}(\infty)) = \left(\nabla q\left(\tilde{w}_1(\infty)\right), ..., \nabla q\left(\tilde{w}_d(\infty)\right)\right)^\top = \sum_{n=1}^N \mathbf{x}^{(n)}\nu^{(n)} \ .$$

We note that $\|\nabla Q_{\mathbf{k}}(\tilde{\mathbf{w}})\| < \infty$ when $\|\tilde{\mathbf{w}}\| < \infty$, and thus by using Lemma 3.1 we get that $\nu^{(n)} < \infty$ for all n. Therefore, we get that gradient flow satisfies the KKT conditions for minimizing this Q, which completes the proof. □

C. Proof of Theorem 6.1

Proof. We start by examining a general multi-neuron fully connected linear network of depth 2, reducing our claim at the end to the case of a network with a single hidden neuron ($m = 1$).

The fully connected linear network of depth 2 is defined as

$$f(\mathbf{x}; \{a_i\}, \{\mathbf{w}_i\}) = \sum_{i=1}^m a_i \mathbf{w}_i^\top \mathbf{x} = \tilde{\mathbf{w}}^\top \mathbf{x} \ ,$$

where $\tilde{\mathbf{w}} \triangleq \sum_{i=1}^m \tilde{\mathbf{w}}_i$, and $\tilde{\mathbf{w}}_i \triangleq a_i \mathbf{w}_i$.

The parameter gradient flow dynamics are given by:

$$\dot{a}_i = -\partial_{a_i}\mathcal{L} = \mathbf{w}_i^\top \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right)$$

$$\dot{\mathbf{w}}_i = -\partial_{\mathbf{w}_i}\mathcal{L} = a_i \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right)$$

$$\frac{d}{dt}\tilde{\mathbf{w}}_i = \dot{a}_i \mathbf{w}_i + a_i \dot{\mathbf{w}}_i = \left(a_i^2 \mathbf{I} + \mathbf{w}_i \mathbf{w}_i^\top\right) \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right) ,$$

where we denote the residual

$$r^{(n)}(t) \triangleq y^{(n)} - \tilde{\mathbf{w}}^\top(t)\mathbf{x}^{(n)} .$$

Using Theorem 2.1 of Du et al. (2018) (stated in Section 6), we can write

$$\frac{d}{dt}\tilde{\mathbf{w}}_i(t) = \left(\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right)\mathbf{I} + \mathbf{w}_i(t)\mathbf{w}_i^\top(t)\right) \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right) , \tag{22}$$

or also

$$\left(\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right)\mathbf{I} + \mathbf{w}_i(t)\mathbf{w}_i^\top(t)\right)^{-1} \frac{d}{dt}\tilde{\mathbf{w}}_i(t) = \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right)$$

where assuming $\delta_i \geq 0$, a non-zero initialization $\tilde{\mathbf{w}}(0) = a(0)\mathbf{w}(0) \neq \mathbf{0}$ and that we converge to zero-loss solution, gives us that the expression $\left(\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right)\mathbf{I} + \mathbf{w}_i(t)\mathbf{w}_i^\top(t)\right)^{-1}$ exists.

Using the Sherman-Morisson Lemma, we have

$$\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right)^{-1} \left(\mathbf{I} - \frac{\mathbf{w}_i(t)\mathbf{w}_i^\top(t)}{\left(\delta_i + 2\|\mathbf{w}_i(t)\|^2\right)}\right) \frac{d}{dt}\tilde{\mathbf{w}}_i = \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right) ,$$

or

$$\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right)^{-1} \left(\mathbf{I} - \frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i^\top(t)}{\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right)\left(\delta_i + 2\|\mathbf{w}_i(t)\|^2\right)}\right) \frac{d}{dt}\tilde{\mathbf{w}}_i = \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right) \tag{23}$$

where we again employed Theorem 2.1 of Du et al. (2018).
Also, since

$$\|\tilde{\mathbf{w}}_i(t)\|^2 = a_i^2(t)\|\mathbf{w}_i(t)\|^2 = \|\mathbf{w}_i(t)\|^2\left(\delta_i + \|\mathbf{w}_i(t)\|^2\right) ,$$

we can express $\mathbf{w}$ as a function of $\tilde{\mathbf{w}}$:

$$\|\mathbf{w}_i(t)\|^2 = \frac{-\delta_i}{2} \pm \sqrt{\frac{\delta_i^2}{4} + \|\tilde{\mathbf{w}}_i(t)\|^2} .$$

Since $\|\mathbf{w}_i(t)\|^2 \geq 0$ we choose the (+) sign and obtain

$$\|\mathbf{w}_i(t)\| = \sqrt{\frac{-\delta_i}{2} + \sqrt{\frac{\delta_i^2}{4} + \|\tilde{\mathbf{w}}_i(t)\|^2}} .$$

Therefore, we can write Eq. 23 as:

$$\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)^{-1}\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i^\top(t)}{2\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}}\right)\frac{d}{dt}\tilde{\mathbf{w}}_i(t)=\sum_{n=1}^{N}\mathbf{x}^{(n)}r^{(n)}(t)\,. \quad (24)$$

We follow the "warped IMD" technique for deriving the implicit bias (presented in detail in Section 5 of the main text) and multiply Eq. 24 by some function $g\left(\tilde{\mathbf{w}}_i(t)\right)$

$$\begin{aligned}&g\left(\tilde{\mathbf{w}}_i(t)\right)\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)^{-1}\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i^\top(t)}{2\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}}\right)\frac{d}{dt}\tilde{\mathbf{w}}_i(t)\\&\quad=\sum_{n=1}^{N}\mathbf{x}^{(n)}g\left(\tilde{\mathbf{w}}_i(t)\right)r^{(n)}(t)\,.\end{aligned}$$

Following the approach in Section 5, we then try and find $q\left(\tilde{\mathbf{w}}_i(t)\right)=\hat{q}\left(\|\tilde{\mathbf{w}}_i(t)\|\right)+\mathbf{z}^\top\tilde{\mathbf{w}}_i(t)$ and $g\left(\tilde{\mathbf{w}}_i(t)\right)$ such that

$$\nabla^2 q\left(\tilde{\mathbf{w}}_i(t)\right)=g\left(\tilde{\mathbf{w}}_i(t)\right)\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)^{-1}\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i^\top(t)}{2\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}}\right)\,, \quad (25)$$

so that then we'll have,

$$\begin{aligned}\nabla^2 q\left(\tilde{\mathbf{w}}_i(t)\right)\frac{d}{dt}\tilde{\mathbf{w}}_i(t)&=\sum_{n=1}^{N}\mathbf{x}^{(n)}g\left(\tilde{\mathbf{w}}_i(t)\right)r^{(n)}(t)\\ \frac{d}{dt}\left(\nabla q\left(\tilde{\mathbf{w}}_i(t)\right)\right)&=\sum_{n=1}^{N}\mathbf{x}^{(n)}g\left(\tilde{\mathbf{w}}_i(t)\right)r^{(n)}(t)\\ \nabla q\left(\tilde{\mathbf{w}}_i(t)\right)-\nabla q\left(\tilde{\mathbf{w}}_i(0)\right)&=\sum_{n=1}^{N}\mathbf{x}^{(n)}\int_0^t g\left(\tilde{\mathbf{w}}_i(t')\right)r^{(n)}(t')dt'\,.\end{aligned}$$

Requiring $\nabla q\left(\tilde{\mathbf{w}}_i(0)\right)=0$, and denoting $\nu_i^{(n)}=\int_0^\infty g\left(\tilde{\mathbf{w}}_i(t')\right)r^{(n)}(t')dt'$, we get the condition:

$$\nabla q\left(\tilde{\mathbf{w}}_i(\infty)\right)=\sum_{n=1}^{N}\mathbf{x}^{(n)}\nu_i^{(n)}\,.$$

To find q we note that:

$$\nabla q\left(\tilde{\mathbf{w}}_i(t)\right)=\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)\frac{\tilde{\mathbf{w}}_i(t)}{\|\tilde{\mathbf{w}}_i(t)\|}+\mathbf{z}$$

and

$$\begin{aligned}\nabla^2 q\left(\tilde{\mathbf{w}}_i(t)\right)&=\left[\hat{q}''\left(\|\tilde{\mathbf{w}}_i(t)\|\right)-\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)\frac{1}{\|\tilde{\mathbf{w}}_i(t)\|}\right]\frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i^\top(t)}{\|\tilde{\mathbf{w}}_i(t)\|^2}+\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)\frac{1}{\|\tilde{\mathbf{w}}_i(t)\|}\mathbf{I}\\&=\frac{\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)}{\|\tilde{\mathbf{w}}_i(t)\|}\left[\mathbf{I}-\left[1-\|\tilde{\mathbf{w}}_i(t)\|\frac{\hat{q}''\left(\|\tilde{\mathbf{w}}_i(t)\|\right)}{\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)}\right]\frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i^\top(t)}{\|\tilde{\mathbf{w}}_i(t)\|^2}\right]\,.\end{aligned}$$

Comparing the form above with the Hessian in Eq. 25 we require

$$g\left(\tilde{\mathbf{w}}_i(t)\right)=\frac{\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)}{\|\tilde{\mathbf{w}}_i(t)\|}\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)$$

and

$$\frac{1}{2\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}} = \frac{1-\|\tilde{\mathbf{w}}_i(t)\|\frac{\hat{q}''(\|\tilde{\mathbf{w}}_i(t)\|)}{\hat{q}'(\|\tilde{\mathbf{w}}_i(t)\|)}}{\|\tilde{\mathbf{w}}_i(t)\|^2}$$

$$\Rightarrow \frac{\hat{q}''\left(\|\tilde{\mathbf{w}}_i(t)\|\right)}{\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)} = \frac{1-\frac{\|\tilde{\mathbf{w}}_i(t)\|^2}{\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i^2}{4}+\|\tilde{\mathbf{w}}_i(t)\|^2}\right)\sqrt{\delta_i^2+4\|\tilde{\mathbf{w}}_i(t)\|^2}}}{\|\tilde{\mathbf{w}}_i(t)\|}$$

$$\Rightarrow \frac{\hat{q}''(x)}{\hat{q}'(x)} = \frac{1-\frac{x^2}{\left(\frac{\delta_i}{2}+\sqrt{\frac{\delta_i{}^2}{4}+x^2}\right)\sqrt{\delta_i^2+4x^2}}}{x}\,.$$

Integrating that we get

$$\log\hat{q}'(x) = \frac{1}{2}\log\left(\sqrt{x^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}\right)+C$$

$$\Rightarrow \hat{q}'(x) = C\sqrt{\sqrt{x^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}$$

$$\Rightarrow \hat{q}(x) = C\frac{\left(x^2-\frac{\delta_i}{2}\left(\frac{\delta_i}{2}+\sqrt{x^2+\frac{\delta_i^2}{4}}\right)\right)\sqrt{\sqrt{x^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}}{x}+C'\,.$$

Therefore,

$$q\left(\tilde{\mathbf{w}}_i(t)\right) = C\frac{\left(\|\tilde{\mathbf{w}}_i(t)\|^2-\frac{\delta_i}{2}\left(\frac{\delta_i}{2}+\sqrt{\|\tilde{\mathbf{w}}_i(t)\|^2+\frac{\delta_i^2}{4}}\right)\right)\sqrt{\sqrt{\|\tilde{\mathbf{w}}_i(t)\|^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}}{\|\tilde{\mathbf{w}}_i(t)\|}+\mathbf{z}^\top\tilde{\mathbf{w}}_i(t)+C'\,.$$

Now, from the condition $\nabla q\left(\tilde{\mathbf{w}}_i(0)\right)=0$ we have

$$\nabla q\left(\tilde{\mathbf{w}}_i(0)\right) = \frac{3}{2}C\frac{\tilde{\mathbf{w}}_i(0)}{\|\tilde{\mathbf{w}}_i(0)\|}\sqrt{\sqrt{\|\tilde{\mathbf{w}}_i(0)\|^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}+\mathbf{z}=0$$

$$\Rightarrow \mathbf{z} = -\frac{3}{2}C\frac{\tilde{\mathbf{w}}_i(0)}{\|\tilde{\mathbf{w}}_i(0)\|}\sqrt{\sqrt{\|\tilde{\mathbf{w}}_i(0)\|^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}\,.$$

We can set $C=1$, $C'=0$ and get

$$\begin{aligned} q\left(\tilde{\mathbf{w}}_i(t)\right) =& \frac{\left(\|\tilde{\mathbf{w}}_i(t)\|^2-\frac{\delta_i}{2}\left(\frac{\delta_i}{2}+\sqrt{\|\tilde{\mathbf{w}}_i(t)\|^2+\frac{\delta_i^2}{4}}\right)\right)\sqrt{\sqrt{\|\tilde{\mathbf{w}}_i(t)\|^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}}{\|\tilde{\mathbf{w}}_i(t)\|} \\ &-\frac{3}{2}\sqrt{\sqrt{\|\tilde{\mathbf{w}}_i(0)\|^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}\frac{\tilde{\mathbf{w}}_i^\top(0)}{\|\tilde{\mathbf{w}}_i(0)\|}\tilde{\mathbf{w}}_i(t)\,. \end{aligned}$$

Finally, for the case of a fully connected network with a single hidden neuron ($m=1$), the condition

$$\nabla q\left(\tilde{\mathbf{w}}_i(\infty)\right) = \sum_{n=1}^{N}\mathbf{x}^{(n)}\nu_i^{(n)}$$

can be written as

$$\nabla q\left(\tilde{\mathbf{w}}(\infty)\right) = \sum_{n=1}^{N}\mathbf{x}^{(n)}\nu^{(n)}$$

in which $\nu^{(n)}$ has no dependency on the index i. Moreover, we note that $\|\nabla q(\tilde{\mathbf{w}})\| < \infty$ when $\|\tilde{\mathbf{w}}\| < \infty$, and thus by using Lemma 3.1 we get that $\nu^{(n)} < \infty$ for all n. And so we got a valid KKT stationarity condition for the q we found above. Therefore, the gradient flow satisfies the KKT conditions for minimizing the q we have found. □

C.1. Validation of the use of the function g as a "Time-Warping"

First, we show that Eq. 24 cannot take the form suggested by Eq. 4 (as in the standard IMD approach described in Section 3):

$$\mathbf{H}(\tilde{\mathbf{w}}(t))\frac{d\tilde{\mathbf{w}}(t)}{dt} = \mathbf{X}\mathbf{r}(t)$$

where $\mathbf{H}(\tilde{\mathbf{w}}(t)) = \nabla^2 Q(\tilde{\mathbf{w}}(t))$ for some Q.

From Eq. 24 we get that $\mathbf{H}(\mathbf{w})$ takes the form

$$\mathbf{H}(\mathbf{w}) = \left(\frac{\delta}{2} + \sqrt{\frac{\delta^2}{4} + \|\mathbf{w}\|^2}\right)^{-1}\left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{2\left(\frac{\delta}{2} + \sqrt{\frac{\delta^2}{4} + \|\mathbf{w}\|^2}\right)\sqrt{\frac{\delta^2}{4} + \|\mathbf{w}\|^2}}\right).$$

Suppose $\mathbf{H}(\mathbf{w})$ is indeed the Hessian of some $Q(\mathbf{w})$, then is must respect the Hessian-map condition (see Eq. 7) for any $\delta \geq 0$. Specifically, for $\delta = 0$ we get

$$\mathbf{H}(\mathbf{w}) = \frac{1}{\|\mathbf{w}\|}\left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{2\|\mathbf{w}\|^2}\right),$$

which does not satisfy the Hessian-map condition

$$\frac{\partial \mathbf{H}_{i,i}(\mathbf{w})}{\partial \mathbf{w}_j} = -\frac{w_j}{\|\mathbf{w}\|^3} + \frac{3}{2}\frac{w_i^2 w_j}{\|\mathbf{w}\|^5} \neq -\frac{w_j}{2\|\mathbf{w}\|^3} + \frac{3}{2}\frac{w_i^2 w_j}{\|\mathbf{w}\|^5} = \frac{\partial \mathbf{H}_{i,j}(\mathbf{w})}{\partial \mathbf{w}_i}.$$

Therefore, Eq. 24 cannot be solved using the standard IMD approach, and requires our suggested "warped IMD" technique (see Section 5).

Second, we write $g\left(\tilde{\mathbf{w}}_i(t)\right)$ explicitly and show it is positive, monotone and bounded.

From Eq. 24 we have

$$g\left(\tilde{\mathbf{w}}_i(t)\right) = \frac{\hat{q}'\left(\|\tilde{\mathbf{w}}_i(t)\|\right)}{\|\tilde{\mathbf{w}}_i(t)\|}\left(\frac{\delta_i}{2} + \sqrt{\frac{\delta_i^2}{4} + \|\tilde{\mathbf{w}}_i(t)\|^2}\right) = \frac{1}{\|\tilde{\mathbf{w}}_i(t)\|}\sqrt{\sqrt{\|\tilde{\mathbf{w}}_i(t)\|^2 + \frac{\delta_i^2}{4}} - \frac{\delta_i}{2}}\left(\frac{\delta_i}{2} + \sqrt{\frac{\delta_i^2}{4} + \|\tilde{\mathbf{w}}_i(t)\|^2}\right).$$

We can see that $g\left(\tilde{\mathbf{w}}_i(t)\right) = \hat{g}\left(\|\tilde{\mathbf{w}}_i(t)\|\right)$ where

$$\hat{g}(x) = \frac{\sqrt{\sqrt{x^2 + \frac{\delta_i^2}{4}} - \frac{\delta_i}{2}}}{x}\left(\frac{\delta_i}{2} + \sqrt{\frac{\delta_i^2}{4} + x^2}\right).$$

We notice that $\hat{g}(x)$ is smooth and positive for $\forall x > 0$, and since $\lim_{x\to 0^+}\hat{g}(x) = \sqrt{\delta_i}$ (see Lemma H.3) it is also bounded for any finite x.

Also, using

$$\hat{g}'(x) = \frac{2\sqrt{x^2 + \frac{\delta_i^2}{4}} - \delta_i}{4\sqrt{x^2 + \frac{\delta_i^2}{4}}\sqrt{\sqrt{x^2 + \frac{\delta_i^2}{4}} - \frac{\delta_i}{2}}}$$

we see that $\hat{g}'(x) > 0$, $\forall x > 0$ and so $\hat{g}(x)$ is monotonically increasing.

C.2. Verification of the Hessian map condition

Finally, we show that $g(\tilde{\mathbf{w}}(t))\mathbf{H}(\tilde{\mathbf{w}}(t))$ does satisfy the Hessian-map condition. We note that this is immediate from the construction of q, but provide it here for completeness.

$$g(\mathbf{w})\mathbf{H}(\mathbf{w}) = \frac{1}{\|\mathbf{w}\|}\sqrt{\sqrt{\|\mathbf{w}\|^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{2\left(\frac{\delta}{2}+\sqrt{\frac{\delta^2}{4}+\|\mathbf{w}\|^2}\right)\sqrt{\frac{\delta^2}{4}+\|\mathbf{w}\|^2}}\right).$$

We denote $f(x)=\frac{1}{x}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}$ and $h(x)=\frac{f(x)}{2\left(\frac{\delta}{2}+\sqrt{\frac{\delta^2}{4}+x^2}\right)\sqrt{\frac{\delta^2}{4}+x^2}}$.

Without loss of generality it is enough to observe the following settings:

$i\neq j\neq k$**:**

$$\frac{\partial\mathbf{H}_{i,j}(\mathbf{w})}{\partial\mathbf{w}_k}=-w_iw_jh'(\|\mathbf{w}\|)\frac{w_k}{\|\mathbf{w}\|}=-w_iw_kh'(\|\mathbf{w}\|)\frac{w_j}{\|\mathbf{w}\|}=\frac{\partial\mathbf{H}_{i,k}(\mathbf{w})}{\partial\mathbf{w}_j}$$

$i=j\neq k$**:**

$$\frac{\partial\mathbf{H}_{i,i}(\mathbf{w})}{\partial\mathbf{w}_k}=f'(\|\mathbf{w}\|)\frac{w_k}{\|\mathbf{w}\|}-w_i^2h'(\|\mathbf{w}\|)\frac{w_k}{\|\mathbf{w}\|}$$

$$\frac{\partial\mathbf{H}_{i,k}(\mathbf{w})}{\partial\mathbf{w}_i}=-w_kh(\|\mathbf{w}\|)-w_iw_kh'(\|\mathbf{w}\|)\frac{w_i}{\|\mathbf{w}\|}=-w_kh(\|\mathbf{w}\|)-w_i^2h'(\|\mathbf{w}\|)\frac{w_k}{\|\mathbf{w}\|}$$

Therefore, if $\forall x$, $\frac{f'(x)}{x}=-h(x)$ we get that $\frac{\partial\mathbf{H}_{i,i}(\mathbf{w})}{\partial\mathbf{w}_k}=\frac{\partial\mathbf{H}_{i,k}(\mathbf{w})}{\partial\mathbf{w}_i}$.

Using the derivative of $f(x)$ we can write:

$$\begin{aligned}
f'(x)&=\frac{1}{2\sqrt{x^2+\frac{\delta^2}{4}}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}-\frac{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{x^2}\\
&=\frac{1}{2\sqrt{x^2+\frac{\delta^2}{4}}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}-\frac{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{\left(\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}\right)\left(\sqrt{x^2+\frac{\delta^2}{4}}+\frac{\delta}{2}\right)}\\
&=\frac{1}{2\sqrt{x^2+\frac{\delta^2}{4}}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}-\frac{1}{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}\left(\sqrt{x^2+\frac{\delta^2}{4}}+\frac{\delta}{2}\right)}\\
&=\frac{\left(\sqrt{x^2+\frac{\delta^2}{4}}+\frac{\delta}{2}\right)-2\sqrt{x^2+\frac{\delta^2}{4}}}{2\sqrt{x^2+\frac{\delta^2}{4}}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}\left(\sqrt{x^2+\frac{\delta^2}{4}}+\frac{\delta}{2}\right)}\\
&=-\frac{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{2\sqrt{x^2+\frac{\delta^2}{4}}\left(\sqrt{x^2+\frac{\delta^2}{4}}+\frac{\delta}{2}\right)}\\
&=-x\cdot h(x)\ ,
\end{aligned}$$

and so $g(\mathbf{w})\mathbf{H}(\mathbf{w})$ respects the Hessian-map condition.

D. Proof of Proposition 6.2

Proof. We recall that the fully connected linear network of depth 2 is defined as

$$f(\mathbf{x};\{a_i\},\{\mathbf{w}_i\}) = \sum_{i=1}^{m} a_i \mathbf{w}_i^\top \mathbf{x} = \tilde{\mathbf{w}}^\top \mathbf{x}\,,$$

where $\tilde{\mathbf{w}} \triangleq \sum_{i=1}^{m} \tilde{\mathbf{w}}_i$, and $\tilde{\mathbf{w}}_i \triangleq a_i \mathbf{w}_i$.

Returning to the dynamics of model parameters (Eq. 22) we have

$$\frac{d}{dt}\tilde{\mathbf{w}}_i(t) = \dot{a}_i \mathbf{w}_i + a_i \dot{\mathbf{w}}_i = \left(a_i^2 \mathbf{I} + \mathbf{w}_i \mathbf{w}_i^\top\right)\left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right)\,.$$

Therefore,

$$\frac{d}{dt}\tilde{\mathbf{w}}(t) = \left(\sum_{i=1}^{m} a_i^2 \mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i \mathbf{w}_i^\top\right)\left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right)$$

$$\left(\sum_{i=1}^{m} a_i^2(t) \mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t) \mathbf{w}_i(t)^\top\right)^{-1} \frac{d}{dt}\tilde{\mathbf{w}}(t) = \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)}\right)\,.$$

We can notice that we can express

$$\sum_{i=1}^{m} a_i^2(t) \mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t) \mathbf{w}_i(t)^\top = \mathbf{A}(t) + \mathbf{U}(t)\mathbf{C}\mathbf{V}(t)$$

where

$$\mathbf{A}(t) = \left(\sum_{i=1}^{m} a_i^2(t)\right)\mathbf{I}_{d\times d}$$
$$\mathbf{C} = \mathbf{I}_{m\times m}$$
$$\mathbf{U}(t) = \mathbf{W}(t) \triangleq [\mathbf{w}_1(t),...,\mathbf{w}_m(t)] \in \mathbb{R}^{d\times m}$$
$$\mathbf{V}(t) = \mathbf{W}(t)^\top = [\mathbf{w}_1(t)^\top;...;\mathbf{w}_m(t)^\top] \in \mathbb{R}^{m\times d}\,.$$

By using the Woodbury matrix identity we can write

$$\left(\sum_{i=1}^{m} a_i^2(t) \mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t) \mathbf{w}_i(t)^\top\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1} =$$
$$= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\left(\sum_{i=1}^{m} a_i^2(t)\right)\mathbf{I} + \mathbf{V}\mathbf{U}\right)^{-1}\mathbf{V}\,.$$

From Theorem 2.2 of Du et al. (2018) (stated in Section 6) we get that

$$\mathbf{a}(t)\cdot\mathbf{a}(t)^\top = \mathbf{W}(t)^\top \mathbf{W}(t) + \mathbf{\Delta}\,,$$

where $\mathbf{\Delta} \in \mathbb{R}^{m\times m}$.

For the case of strict balanced initialization we have $\mathbf{\Delta} = 0$, and therefore

$$\left(\left(\sum_{i=1}^{m} a_i^2(t)\right)\mathbf{I} + \mathbf{V}(t)\mathbf{U}(t)\right)^{-1} = \left(\left(\sum_{i=1}^{m} a_i^2(t)\right)\mathbf{I} + \mathbf{W}(t)^\top \mathbf{W}(t)\right)^{-1} =$$
$$= \left(\left(\sum_{i=1}^{m} a_i^2(t)\right)\mathbf{I} + \mathbf{a}(t)\mathbf{a}(t)^\top\right)^{-1}$$
$$= \left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1}\mathbf{I} - \frac{\mathbf{a}(t)\mathbf{a}(t)^\top}{2\left(\sum_{i=1}^{m} a_i^2(t)\right)^2}\,,$$

where in the last transition we used the Sherman-Morrison lemma. It follows that

$$\left(\sum_{i=1}^{m} a_i^2(t)\mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t)\mathbf{w}_i(t)^\top\right)^{-1} =$$
$$= \left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1} \left(\mathbf{I} - \mathbf{W}(t)\left(\left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1}\mathbf{I} - \frac{\mathbf{a}(t)\mathbf{a}(t)^\top}{2\left(\sum_{i=1}^{m} a_i^2(t)\right)^2}\right)\mathbf{W}(t)^\top\right) .$$

We continue and write

$$\left(\sum_{i=1}^{m} a_i^2(t)\mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t)\mathbf{w}_i(t)^\top\right)^{-1} =$$
$$= \left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1} \left(\mathbf{I} - \left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1}\mathbf{W}(t)\mathbf{W}(t)^\top + \frac{1}{2}\left(\sum_{i=1}^{m} a_i^2(t)\right)^{-2}\left(\mathbf{W}(t)\mathbf{W}(t)^\top\right)^2\right) .$$

Using Theorem 2.1 of Du et al. (2018) (stated in Section 6), we know that

$$a_i(t)^2 = \|\mathbf{w}_i(t)\|^2 .$$

Therefore,

$$\|\tilde{\mathbf{w}}_i(t)\| = |a_i(t)|\|\mathbf{w}_i(t)\| = a_i(t)^2$$

and

$$\sum_{i=1}^{m} a_i^2(t) = \sum_{i=1}^{m} \|\tilde{\mathbf{w}}_i(t)\| .$$

So, we can write

$$\left(\sum_{i=1}^{m} a_i^2(t)\mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t)\mathbf{w}_i(t)^\top\right)^{-1} =$$
$$= \left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1} \left(\mathbf{I} - \left(\sum_{i=1}^{m} a_i^2(t)\right)^{-1}\left(\sum_{i=1}^{m} \mathbf{w}_i(t)\mathbf{w}_i(t)^\top\right) + \frac{1}{2}\left(\sum_{i=1}^{m} a_i^2(t)\right)^{-2}\left(\sum_{i=1}^{m} \mathbf{w}_i(t)\mathbf{w}_i(t)^\top\right)^2\right) =$$
$$= \left(\sum_{i=1}^{m} \|\tilde{\mathbf{w}}_i(t)\|\right)^{-1} \left(\mathbf{I} - \left(\sum_{i=1}^{m} \|\tilde{\mathbf{w}}_i(t)\|\right)^{-1}\left(\sum_{i=1}^{m} \frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i(t)^\top}{\|\tilde{\mathbf{w}}_i(t)\|}\right) + \frac{1}{2}\left(\sum_{i=1}^{m} \|\tilde{\mathbf{w}}_i(t)\|\right)^{-2}\left(\sum_{i=1}^{m} \frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i(t)^\top}{\|\tilde{\mathbf{w}}_i(t)\|}\right)^2\right) .$$

Now, since

$$\mathbf{a}(t)\mathbf{a}(t)^\top = \mathbf{W}(t)^\top\mathbf{W}(t) ,$$

we can say that $\mathbf{W}(t)^\top\mathbf{W}(t)$ is a rank one matrix, and therefore also $\mathbf{W}(t)$, and also $\tilde{\mathbf{W}}(t)$.
Therefore, all $\tilde{\mathbf{w}}_i$ are equal up to a multiplicative factor,

$$\tilde{\mathbf{w}}_i(t) = c_i(t)\tilde{\mathbf{w}}(t)$$

where from definition

$$\sum_{i=1}^{m} c_i(t) = 1 .$$

Therefore,

$$\|\tilde{\mathbf{w}}_i(t)\| = |c_i(t)|\|\tilde{\mathbf{w}}(t)\|$$
$$\Rightarrow \sum_{i=1}^{m} \|\tilde{\mathbf{w}}_i(t)\| = \left(\sum_{i=1}^{m} |c_i(t)|\right)\|\tilde{\mathbf{w}}(t)\|$$

$$\Rightarrow \sum_{i=1}^{m} \frac{\tilde{\mathbf{w}}_i(t)\tilde{\mathbf{w}}_i(t)^\top}{\|\tilde{\mathbf{w}}_i(t)\|} = \left(\sum_{i=1}^{m} |c_i(t)|\right) \frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|} ,$$

giving us

$$\begin{aligned}
&\left(\sum_{i=1}^{m} a_i^2(t)\mathbf{I} + \sum_{i=1}^{m} \mathbf{w}_i(t)\mathbf{w}_i(t)^\top\right)^{-1} \frac{d}{dt}\tilde{\mathbf{w}}(t) = \\
&= \frac{1}{\left(\sum_{i=1}^{m} |c_i(t)|\right)} \frac{1}{|\|\tilde{\mathbf{w}}(t)\|} \left(\mathbf{I} - \frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2} + \frac{1}{2}\left(\frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2}\right)^2\right) \frac{d}{dt}\tilde{\mathbf{w}}(t) = \sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)} \\
&\Rightarrow \frac{1}{\left(\sum_{i=1}^{m} |c_i(t)|\right)} \frac{1}{|\|\tilde{\mathbf{w}}(t)\|} \left(\mathbf{I} - \frac{1}{2}\frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2}\right) \frac{d}{dt}\tilde{\mathbf{w}}(t) = \sum_{n=1}^{N} \mathbf{x}^{(n)} r^{(n)} ,
\end{aligned}$$

where in the last transition we used

$$\left(\frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2}\right)^2 = \frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^4} = \frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2} .$$

We follow the "warped IMD" technique (presented in detail in Section 5) and multiply the equation by some function $g\left(\tilde{\mathbf{w}}_i(t)\right)$

$$\frac{g\left(\tilde{\mathbf{w}}(t)\right)}{\left(\sum_{i=1}^{m} |c_i(t)|\right)} \frac{1}{|\|\tilde{\mathbf{w}}(t)\|} \left(\mathbf{I} - \frac{1}{2}\frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2}\right) \frac{d}{dt}\tilde{\mathbf{w}}(t) = \left(\sum_{n=1}^{N} \mathbf{x}^{(n)} g\left(\tilde{\mathbf{w}}(t)\right) r^{(n)}\right) .$$

Following the approach in Section 5, we then try and find $q\left(\tilde{\mathbf{w}}_i(t)\right) = \hat{q}\left(\|\tilde{\mathbf{w}}_i(t)\|\right) + \mathbf{z}^\top \tilde{\mathbf{w}}_i(t)$ and $g\left(\tilde{\mathbf{w}}_i(t)\right)$ such that

$$\nabla^2 q\left(\tilde{\mathbf{w}}(t)\right) = \frac{g\left(\tilde{\mathbf{w}}(t)\right)}{\left(\sum_{i=1}^{m} |c_i(t)|\right)} \frac{1}{|\|\tilde{\mathbf{w}}(t)\|} \left(\mathbf{I} - \frac{1}{2}\frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}(t)^\top}{\|\tilde{\mathbf{w}}(t)\|^2}\right) , \tag{26}$$

so that then we'll have

$$\nabla^2 q\left(\tilde{\mathbf{w}}(t)\right) \frac{d}{dt}\tilde{\mathbf{w}}(t) = \sum_{n=1}^{N} \mathbf{x}^{(n)} g\left(\tilde{\mathbf{w}}(t)\right) r^{(n)}(t)$$

$$\frac{d}{dt}\left(\nabla q\left(\tilde{\mathbf{w}}(t)\right)\right) = \sum_{n=1}^{N} \mathbf{x}^{(n)} g\left(\tilde{\mathbf{w}}(t)\right) r^{(n)}(t)$$

$$\nabla q\left(\tilde{\mathbf{w}}(t)\right) - \nabla q\left(\tilde{\mathbf{w}}(0)\right) = \sum_{n=1}^{N} \mathbf{x}^{(n)} \int_0^t g\left(\tilde{\mathbf{w}}(t')\right) r^{(n)}(t')dt' .$$

Assuming $\nabla q\left(\tilde{\mathbf{w}}(0)\right) = 0$, and denoting $\nu^{(n)} = \int_0^\infty g\left(\tilde{\mathbf{w}}(t')\right) r^{(n)}(t')dt'$, we get the condition

$$\nabla q\left(\tilde{\mathbf{w}}(\infty)\right) = \sum_{n=1}^{N} \mathbf{x}^{(n)} \nu^{(n)} .$$

To find q we note that

$$\nabla q\left(\tilde{\mathbf{w}}(t)\right) = \hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right) \frac{\tilde{\mathbf{w}}(t)}{\|\tilde{\mathbf{w}}(t)\|} + \mathbf{z}$$

and

$$\begin{aligned}
\nabla^2 q\left(\tilde{\mathbf{w}}(t)\right) &= \left[\hat{q}''\left(\|\tilde{\mathbf{w}}(t)\|\right) - \hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right) \frac{1}{\|\tilde{\mathbf{w}}(t)\|}\right] \frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}^\top(t)}{\|\tilde{\mathbf{w}}(t)\|^2} + \hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right) \frac{1}{\|\tilde{\mathbf{w}}(t)\|}\mathbf{I} \\
&= \frac{\hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right)}{\|\tilde{\mathbf{w}}(t)\|} \left[\mathbf{I} - \left[1 - \|\tilde{\mathbf{w}}(t)\| \frac{\hat{q}''\left(\|\tilde{\mathbf{w}}(t)\|\right)}{\hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right)}\right] \frac{\tilde{\mathbf{w}}(t)\tilde{\mathbf{w}}^\top(t)}{\|\tilde{\mathbf{w}}(t)\|^2}\right] .
\end{aligned}$$

Comparing the form above with the Hessian in Eq. 26 we require

$$\frac{g\left(\tilde{\mathbf{w}}(t)\right)}{\left(\sum_{i=1}^{m}|c_i(t)|\right)}=\hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right)\ ,$$

and

$$1-\|\tilde{\mathbf{w}}(t)\|\frac{\hat{q}''\left(\|\tilde{\mathbf{w}}(t)\|\right)}{\hat{q}'\left(\|\tilde{\mathbf{w}}(t)\|\right)}=\frac{1}{2}$$

$$\Rightarrow\frac{\hat{q}''\left(x\right)}{\hat{q}'\left(x\right)}=\frac{1}{2x}$$

$$\log\hat{q}'\left(x\right)=\frac{1}{2}\ln x+C$$

$$\hat{q}'\left(x\right)=C\sqrt{x}\ .$$

Therefore,

$$q\left(\tilde{\mathbf{w}}(t)\right)=C\left\|\tilde{\mathbf{w}}(t)\right\|^{3/2}+\mathbf{z}^{\top}\tilde{\mathbf{w}}(t)+C'\ ,$$

and using the condition $\nabla q\left(\tilde{\mathbf{w}}(0)\right)=0$ we get

$$q\left(\tilde{\mathbf{w}}(t)\right)=C\left\|\tilde{\mathbf{w}}(t)\right\|^{3/2}-C\frac{3}{2}\left\|\tilde{\mathbf{w}}(0)\right\|^{-1/2}\tilde{\mathbf{w}}(0)^{\top}\tilde{\mathbf{w}}(t)+C'\ .$$

We can set $C=1$, $C'=0$ and get

$$q\left(\tilde{\mathbf{w}}(t)\right)=\left\|\tilde{\mathbf{w}}(t)\right\|^{3/2}-\frac{3}{2}\left\|\tilde{\mathbf{w}}(0)\right\|^{-1/2}\tilde{\mathbf{w}}(0)^{\top}\tilde{\mathbf{w}}(t)\ .$$

We note that $\|\nabla q(\tilde{\mathbf{w}})\|<\infty$ when $\|\tilde{\mathbf{w}}\|<\infty$, and thus by using Lemma 3.1 we get that $\hat{\nu}^{(n)}<\infty$ for all n. Therefore, gradient flow satisfies the KKT conditions for minimizing this q. □

E. Proof of Theorem 6.3

We recall the proof of Theorem 6.1 given in Appendix C.

The form of the q function described in the proof is $q\left(\tilde{\mathbf{w}}_i(t)\right)=\hat{q}\left(\|\tilde{\mathbf{w}}_i(t)\|\right)+\mathbf{z}^{\top}\tilde{\mathbf{w}}_i(t)$, where

$$\mathbf{z}=-\frac{3}{2}\sqrt{\sqrt{\left\|\tilde{\mathbf{w}}(0)\right\|^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}\frac{\tilde{\mathbf{w}}(0)}{\|\tilde{\mathbf{w}}(0)\|}\ .$$

Under the limit $\|\tilde{\mathbf{w}}_i(0)\|\to 0$ we can see that $\|\mathbf{z}\|\to 0$.

When the linear term captured by $\mathbf{z}$ in the q function is equal to zero, we have

$$\nabla q\left(\tilde{\mathbf{w}}_i(\infty)\right)=\hat{q}'\left(\|\tilde{\mathbf{w}}_i(\infty)\|\right)\frac{\tilde{\mathbf{w}}_i(\infty)}{\|\tilde{\mathbf{w}}_i(\infty)\|}=\sum_n\mathbf{x}^{(n)}\nu_i^{(n)}\ .$$

Defining $\hat{\nu}_i^{(n)}=\frac{\nu_i^{(n)}\|\tilde{\mathbf{w}}_i(\infty)\|}{\hat{q}'(\|\tilde{\mathbf{w}}_i(\infty)\|)}$ we get

$$\tilde{\mathbf{w}}_i(\infty)=\sum_n\mathbf{x}^{(n)}\hat{\nu}_i^{(n)}\ .$$

We notice that

$$\hat{q}'\left(x\right)=\sqrt{\sqrt{x^2+\frac{\delta_i^2}{4}}-\frac{\delta_i}{2}}$$

Using the linear predictor definition of $\tilde{\mathbf{w}}(\infty)=\sum_i\tilde{\mathbf{w}}_i(\infty)$, denoting $\hat{\nu}^{(n)}=\sum_i\hat{\nu}_i^{(n)}$ and summing over i gives

$$\tilde{\mathbf{w}}(\infty) = \sum_n \mathbf{x}^{(n)} \hat{\nu}^{(n)}$$

We note that $\|\nabla q(\tilde{\mathbf{w}})\| < \infty$ when $\|\tilde{\mathbf{w}}\| < \infty$, and thus by using Lemma 3.1 we get that $\hat{\nu}^{(n)} < \infty$ for all n. Therefore, the above is a valid KKT stationarity condition of the form $\nabla q\left(\tilde{\mathbf{w}}(\infty)\right) = \sum_n \mathbf{x}^{(n)} \hat{\nu}^{(n)}$ with $\nabla q\left(\mathbf{w}\right) = \mathbf{w}$. Hence, gradient flow satisfies the KKT conditions for minimizing this q.

It follows that for a multi-neuron fully connected network with non-zero infinitesimal initialization,

$$\tilde{\mathbf{w}}(\infty) = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\|^2 \;\; \text{s.t. } \mathbf{X}^\top \mathbf{w} = \mathbf{y}$$

which is equivalent to

$$\tilde{\mathbf{w}}(\infty) = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\| \;\; \text{s.t. } \mathbf{X}^\top \mathbf{w} = \mathbf{y} \, .$$

F. Characterization of the Implicit Bias Captured in Theorem 4.1

In this Appendix we provide a detailed characterization of the implicit bias for a diagonal linear network as described in Theorem 4.1,

$$\tilde{\mathbf{w}}(\infty) = \arg\min_{\mathbf{w}} Q_{\boldsymbol{k}}(\mathbf{w}) \quad \text{s.t. } \mathbf{X}^\top \mathbf{w} = \mathbf{y}$$

where

$$Q_{\boldsymbol{k}}\left(\mathbf{w}\right) = \sum_{i=1}^{d} q_{k_i}\left(w_i\right) \;,$$

$$q_k\left(x\right) = \frac{\sqrt{k}}{4}\left[1 - \sqrt{1 + \frac{4x^2}{k}} + \frac{2x}{\sqrt{k}} \operatorname{arcsinh}\left(\frac{2x}{\sqrt{k}}\right)\right]$$

and

$$\sqrt{k_i} = \frac{4\alpha_i\left(1 + s_i^2\right)}{1 - s_i^2} \, .$$

For simplicity, we next assume $\alpha_i = \alpha, \;\; s_i = s \;\; \forall i \in [d]$.

We can notice that for $k \to \infty$, i.e. $\frac{\alpha}{1-s^2} \to \infty$ we get that:

$$q_k\left(w_i\right) \xrightarrow{k\to\infty} \frac{w_i^2}{\sqrt{k}} = \frac{1}{2\left(u_{+,i}^2\left(0\right) + v_{+,i}^2\left(0\right)\right)} w_i^2$$
$$\Rightarrow Q_{\boldsymbol{k}}\left(\mathbf{w}\right) = \sum_{i=1}^{d} q_k\left(w_i\right) = \sum_{i=1}^{d} \frac{1}{2\left(u_{+,i}^2\left(0\right) + v_{+,i}^2\left(0\right)\right)} w_i^2 \, .$$

Calculating the tangent kernel at the initialization we get

$$\begin{aligned} K(\mathbf{x}_1, \mathbf{x}_2) =& \langle \nabla f\left(\mathbf{x}_1\right), \nabla f\left(\mathbf{x}_2\right) \rangle \\ =& \langle \left[\mathbf{x}_1 \circ \mathbf{u}_+\left(0\right), \mathbf{x}_1 \circ \mathbf{v}_+\left(0\right), -\mathbf{x}_1 \circ \mathbf{u}_-\left(0\right), -\mathbf{x}_1 \circ \mathbf{v}_-\left(0\right)\right], \\ & \left[\mathbf{x}_2 \circ \mathbf{u}_+\left(0\right), \mathbf{x}_2 \circ \mathbf{v}_+\left(0\right), -\mathbf{x}_2 \circ \mathbf{u}_-\left(0\right), -\mathbf{x}_2 \circ \mathbf{v}_-\left(0\right)\right] \rangle \\ =& \mathbf{x}_1^\top \operatorname{diag}\left(\mathbf{u}_+^2\left(0\right) + \mathbf{v}_+^2\left(0\right) + \mathbf{u}_-^2\left(0\right) + \mathbf{v}_-^2\left(0\right)\right) \mathbf{x}_2 \, . \end{aligned}$$

For the case of unbiased initialization $\left(u_{+,i}\left(0\right) = u_{-,i}\left(0\right), v_{+,i}\left(0\right) = v_{-,i}\left(0\right)\right)$ we have

$$K(\mathbf{x}_1, \mathbf{x}_2) = 2\mathbf{x}_1^\top \operatorname{diag}\left(\mathbf{u}_+^2\left(0\right) + \mathbf{v}_+^2\left(0\right)\right) \mathbf{x}_2 \, .$$

Therefore, using Lemma H.4, we can see that $Q_{\boldsymbol{k}}(\mathbf{w})$ is the RKHS norm with respect to the NTK at initialization. Therefore, $k \to \infty$ indeed describes the NTK regime.

For $k \to 0$, i.e. $\frac{\alpha}{1-s^2} \to 0$ we get that:

$$\begin{aligned}
q_k\left(w_i\right) &= \frac{\sqrt{k}}{4}\left[1-\sqrt{1+\frac{4w_i^2}{k}}+\frac{2w_i}{\sqrt{k}}\text{arcsinh}\left(\frac{2w_i}{\sqrt{k}}\right)\right] \\
&= \frac{\sqrt{k}}{4}-\sqrt{\frac{k}{16}+\frac{w_i^2}{4}}+\frac{w_i}{2}\text{arcsinh}\left(\frac{2w_i}{\sqrt{k}}\right) \\
&\xrightarrow{k\to 0} \frac{\left|w_i\right|}{2}+\frac{\left|w_i\right|}{2}\log\left(\frac{4\left|w_i\right|}{\sqrt{k}}\right) \\
&= \frac{1}{2}\left[-\left|w_i\right|+\left|w_i\right|\log\left(\frac{4\left|w_i\right|}{\sqrt{k}}\right)\right] \\
&= \frac{1}{2}\left[\left|w_i\right|\log\left(\frac{1}{\sqrt{k}}\right)+\left|w_i\right|\left(\log\left(4\left|w_i\right|\right)-1\right)\right]
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow \frac{q_k\left(w_i\right)}{\frac{1}{2}\log\left(\frac{1}{\sqrt{k}}\right)} \to \left|w_i\right|+\frac{\left|w_i\right|\left(\log\left(4\left|w_i\right|\right)-1\right)}{\log\left(\frac{1}{\sqrt{k}}\right)} \\
&= \left|w_i\right|+O\left(\frac{1}{\log\left(\frac{1}{\sqrt{k}}\right)}\right) \to \left|w_i\right|
\end{aligned}$$

Therefore,

$$Q_{\boldsymbol{k}}\left(\mathbf{w}\right)=\sum_{i=1}^{d}\left|w_i\right|=\left\|\mathbf{w}\right\|_1$$

and $k \to 0$ describes the rich regime (Woodworth et al., 2020).

G. Characterization of the Implicit Bias Captured in Theorem 6.1

In this Appendix we provide a detailed characterization of the implicit bias for a two-layer fully connected neural network with a single hidden neuron ($m = 1$) described in Theorem 6.1,

$$\tilde{\mathbf{w}}(\infty) = \arg\min_{\mathbf{w}} q(\mathbf{w}) \quad \text{s.t. } \mathbf{X}^\top \mathbf{w} = \mathbf{y}$$

$$q\left(\tilde{\mathbf{w}}\right) = \hat{q}\left(\left\|\tilde{\mathbf{w}}\right\|\right) + \mathbf{z}^\top \tilde{\mathbf{w}}\ ,$$

where

$$\hat{q}\left(x\right) = \frac{\left(x^2-\frac{\delta}{2}\left(\frac{\delta}{2}+\sqrt{x^2+\frac{\delta^2}{4}}\right)\right)\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{x}$$

$$\mathbf{z} = -\frac{3}{2}\sqrt{\sqrt{\left\|\tilde{\mathbf{w}}\left(0\right)\right\|^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}\,\frac{\tilde{\mathbf{w}}\left(0\right)}{\left\|\tilde{\mathbf{w}}\left(0\right)\right\|}\ .$$

Note that for the sake of simplicity the notations above are an abbreviated version of those found Theorem 6.1.

We will employ the initialization orientation, defined as $\mathbf{u} = \frac{\mathbf{w}(0)}{\|\mathbf{w}(0)\|}$, and the initialization scale, $\left\|\tilde{\mathbf{w}}\left(0\right)\right\| = \alpha$.

G.1. The case $\alpha \to 0$ for any $0 \le s < 1$

Note that from Lemma H.2 (part 2) we have

$$\|\mathbf{z}\| = \frac{3}{2}\sqrt{\sqrt{\|\tilde{\mathbf{w}}(0)\|^2 + \frac{\delta^2}{4}} - \frac{\delta}{2}} = \frac{3}{2}\sqrt{\sqrt{\alpha^2 + \frac{\delta^2}{4}} - \frac{\delta}{2}} = \frac{3}{2}\sqrt{\alpha\frac{1-s}{1+s}}\,,$$

and thus for any $0 \le s < 1$ when $\alpha \to 0$ we get that $\|\mathbf{z}\| \to 0$. It follows that $q_\delta(\tilde{\mathbf{w}}) = \hat{q}(\|\tilde{\mathbf{w}}\|)$ and since $\hat{q}(x)$ is a monotonically increasing function (for any δ) we get the ℓ_2 implicit bias,

$$\tilde{\mathbf{w}}(\infty) = \arg\min_{\tilde{\mathbf{w}}}(q_\delta(\tilde{\mathbf{w}})) = \arg\min_{\tilde{\mathbf{w}}}(\hat{q}(\|\tilde{\mathbf{w}}\|)) = \arg\min_{\tilde{\mathbf{w}}}\|\tilde{\mathbf{w}}\|\ .$$

We call this regime the *Anti-NTK* regime.

G.2. Other special cases

Here we analyze the Taylor expansion of $q(\tilde{\mathbf{w}})$ around $\tilde{\mathbf{w}}(0)$. To this end, we know that

$$\nabla^2 q(\tilde{\mathbf{w}}) = \frac{\hat{q}'(\|\tilde{\mathbf{w}}\|)}{\|\tilde{\mathbf{w}}\|}\left(\mathbf{I} - \frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{2\left(\frac{\delta}{2} + \sqrt{\frac{\delta^2}{4} + \|\tilde{\mathbf{w}}\|^2}\right)\sqrt{\frac{\delta^2}{4} + \|\tilde{\mathbf{w}}\|^2}}\right)\,,$$

and thus the third-order term is order of $\frac{d}{dx}\frac{\hat{q}'(x)}{x}(\|\tilde{\mathbf{w}}(0)\|)$. Since we know that $\nabla q(\tilde{\mathbf{w}}(0)) = 0$ we can write the Taylor expansion as follows

$$q(\tilde{\mathbf{w}}) = q(\tilde{\mathbf{w}}(0)) + \frac{1}{2}(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0))^\top \nabla^2 q(\tilde{\mathbf{w}}(0))(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0)) + O\left(\frac{d}{dx}\frac{\hat{q}'(x)}{x}(\|\tilde{\mathbf{w}}(0)\|)\right)\ .$$

By using Lemma H.2 and

$$\hat{q}'(x) = \sqrt{\sqrt{x^2 + \frac{\delta^2}{4}} - \frac{\delta}{2}}$$

we calculate

$$\begin{aligned}
\nabla^2 q(\tilde{\mathbf{w}}(0)) &= \frac{\hat{q}'(\|\tilde{\mathbf{w}}(0)\|)}{\|\tilde{\mathbf{w}}(0)\|}\left(\mathbf{I} - \frac{\tilde{\mathbf{w}}(0)\tilde{\mathbf{w}}(0)^\top}{\left(\frac{\delta}{2} + \sqrt{\frac{\delta^2}{4} + \|\tilde{\mathbf{w}}(0)\|^2}\right)\sqrt{\delta^2 + 4\|\tilde{\mathbf{w}}(0)\|^2}}\right) \\
&= \frac{\sqrt{\sqrt{\|\tilde{\mathbf{w}}(0)\|^2 + \frac{\delta^2}{4}} - \frac{\delta}{2}}}{\|\tilde{\mathbf{w}}(0)\|}\left(\mathbf{I} - \frac{\tilde{\mathbf{w}}(0)\tilde{\mathbf{w}}(0)^\top}{2\left(\frac{\delta}{2} + \sqrt{\frac{\delta^2}{4} + \|\tilde{\mathbf{w}}(0)\|^2}\right)\sqrt{\frac{\delta^2}{4} + \|\tilde{\mathbf{w}}(0)\|^2}}\right) \\
&= \frac{\sqrt{\sqrt{\alpha^2 + \frac{\delta^2}{4}} - \frac{\delta}{2}}}{\alpha}\left(\mathbf{I} - \frac{\alpha^2\mathbf{u}\mathbf{u}^\top}{2\left(\frac{\delta}{2} + \sqrt{\frac{\delta^2}{4} + \alpha^2}\right)\sqrt{\frac{\delta^2}{4} + \alpha^2}}\right) \\
&= \sqrt{\frac{1-s}{\alpha}}\sqrt{\frac{1}{1+s}}\left(\mathbf{I} - \frac{(1-s)^2}{2(1+s^2)}\mathbf{u}\mathbf{u}^\top\right)\ .
\end{aligned}$$

Also, by using

$$\hat{q}''(x) = \frac{x}{2\sqrt{x^2 + \frac{\delta^2}{4}}\sqrt{\sqrt{x^2 + \frac{\delta^2}{4}} - \frac{\delta}{2}}}$$

we have that

$$\begin{aligned}\frac{d}{dx}\frac{\hat{q}'(x)}{x} &= \frac{\hat{q}''(x)\,x-\hat{q}'(x)}{x^2}\\ &= \frac{\frac{x^2}{2\sqrt{x^2+\frac{\delta^2}{4}}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}-\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{x^2}\\ &= \frac{1}{2\sqrt{x^2+\frac{\delta^2}{4}}\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}-\frac{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{x^2}\,,\end{aligned}$$

and thus, using Lemma H.2 we get

$$\begin{aligned}\frac{d}{dx}\frac{\hat{q}'(x)}{x}\left(\|\tilde{\mathbf{w}}(0)\|\right) &= \frac{1}{2\sqrt{\alpha^2+\frac{\delta^2}{4}}\sqrt{\sqrt{\alpha^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}-\frac{\sqrt{\sqrt{\alpha^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{\alpha^2}\\ &= -\frac{(1-s)^{2.5}}{\alpha^{1.5}}\left(\frac{1}{2\left(1+s^2\right)\sqrt{1+s}}\right)\,.\end{aligned}$$

Therefore, the Taylor expansion is

$$q\left(\tilde{\mathbf{w}}\right)=q\left(\tilde{\mathbf{w}}\left(0\right)\right)+\frac{1}{2}\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}\left(0\right)\right)^{\top}\left[\sqrt{\frac{1-s}{\alpha}}\sqrt{\frac{1}{1+s}}\left(\mathbf{I}-\frac{(1-s)^2}{2\left(1+s^2\right)}\mathbf{u}\mathbf{u}^{\top}\right)\right]\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}\left(0\right)\right)+O\left(\frac{(1-s)^{2.5}}{\alpha^{1.5}}\left(\frac{1}{2\left(1+s^2\right)\sqrt{1+s}}\right)\right)\,.$$

We are interested in cases where the higher order terms vanish. Since $0\le s<1$, we only need to require

$$\frac{(1-s)^{2.5}}{\alpha^{1.5}}\ll\sqrt{\frac{1-s}{\alpha}}$$

$$\Rightarrow\frac{(1-s)^2}{\alpha}\ll 1\,. \tag{27}$$

In follows that when $\frac{(1-s)^2}{\alpha}\ll 1$ we can approximate

$$q\left(\tilde{\mathbf{w}}\right)\approx q\left(\tilde{\mathbf{w}}\left(0\right)\right)+\frac{1}{2}\sqrt{\frac{1-s}{\alpha}}\sqrt{\frac{1}{1+s}}\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}\left(0\right)\right)^{\top}\left(\mathbf{I}-\frac{(1-s)^2}{2\left(1+s^2\right)}\mathbf{u}\mathbf{u}^{\top}\right)\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}\left(0\right)\right)\,.$$

In this case, minimizing $q\left(\tilde{\mathbf{w}}\right)$ boils down to minimizing the squared Mahalanobis norm

$$\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}\left(0\right)\right)^{\top}\mathbf{B}\left(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}\left(0\right)\right)$$

where

$$\mathbf{B}=\mathbf{I}-\frac{(1-s)^2}{2\left(1+s^2\right)}\mathbf{u}\mathbf{u}^{\top}\,. \tag{28}$$

Note that $\mathbf{B}^{-1}$ is related to the NTK at initialization, since it is easy to verify that

$$\mathbf{B}^{-1}=\frac{1}{a(0)^2}\left(a(0)^2\mathbf{I}+\mathbf{w}(0)\mathbf{w}(0)^{\top}\right)\,,$$

and the NTK at initialization is given by

$$K(\mathbf{x},\mathbf{x}')=\mathbf{x}^{\top}\left(a(0)^2\mathbf{I}+\mathbf{w}(0)\mathbf{w}(0)^{\top}\right)\mathbf{x}'=a(0)^2\left(\mathbf{x}^{\top}\mathbf{B}^{-1}\mathbf{x}'\right)\,.$$

More specifically, using Lemma H.4, we can see that $q\left(\tilde{\mathbf{w}}\right)$ is the RKHS norm with respect to the NTK at initialization.

Next, we discuss the cases when condition (27) holds.

G.2.1. THE CASE $\alpha \to \infty$ FOR ANY $0 \leq s < 1$

In this case (27) holds and thus the implicit bias is given by

$$\tilde{\mathbf{w}}(\infty) = \arg\min_{\tilde{\mathbf{w}}} \left(q_\delta(\tilde{\mathbf{w}})\right) = \arg\min_{\tilde{\mathbf{w}}} \left((\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0))^\top \mathbf{B} (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0)) \right) ,$$

where $\mathbf{B}$ defined in (28).

G.2.2. THE CASE $s \to 1$ FOR ANY $\alpha > 0$

In this case (27) also holds and thus the implicit bias is given by

$$\tilde{\mathbf{w}}(\infty) = \arg\min_{\tilde{\mathbf{w}}} \left(q_\delta(\tilde{\mathbf{w}})\right) = \arg\min_{\tilde{\mathbf{w}}} \left((\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0))^\top \mathbf{B} (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0)) \right) ,$$

where $\mathbf{B}$ defined in (28). Since $s \to 1$ we get that $\mathbf{B} \to \mathbf{I}$ and thus

$$\tilde{\mathbf{w}}(\infty) = \arg\min_{\tilde{\mathbf{w}}} \left(\|\tilde{\mathbf{w}} - \tilde{\mathbf{w}}(0)\|\right) .$$

H. Auxiliary Lemmas

Lemma H.1. $\delta = a^2(0) - \|\mathbf{w}(0)\|^2 = \frac{4\alpha s}{1-s^2}$.

Proof. By the notation

$$\alpha = |a(0)| \cdot \|\mathbf{w}(0)\|$$
$$s = \frac{|a(0)| - \|\mathbf{w}(0)\|}{|a(0)| + \|\mathbf{w}(0)\|}$$

we get

$$1 - s^2 = \frac{4|a(0)|\|\mathbf{w}(0)\|}{(|a(0)| + \|\mathbf{w}(0)\|)^2}$$

and

$$\begin{aligned}\frac{4\alpha s}{1-s^2} &= 4\alpha \frac{|a(0)| - \|\mathbf{w}(0)\|}{|a(0)| + \|\mathbf{w}(0)\|} \frac{(|a(0)| + \|\mathbf{w}(0)\|)^2}{4\alpha} \\ &= a^2(0) - \|\mathbf{w}(0)\|^2 = \delta .\end{aligned}$$

□

Lemma H.2. *The initialization scale α, initialization shape s and the balancedness factor δ satisfy:*

1.
$$\sqrt{\alpha^2 + \frac{\delta^2}{4}} = \frac{\alpha\left(1+s^2\right)}{1-s^2}$$

2.
$$\sqrt{\alpha^2 + \frac{\delta^2}{4}} - \frac{\delta}{2} = \alpha \frac{1-s}{1+s}$$

3.
$$\sqrt{\alpha^2 + \frac{\delta^2}{4}} + \frac{\delta}{2} = \alpha \frac{1+s}{1-s}$$

Proof. 1. Using Lemma H.1 we get

$$\sqrt{\alpha^2 + \frac{\delta^2}{4}} = \sqrt{\alpha^2 + \frac{4\alpha^2 s^2}{(1-s^2)^2}} = \frac{\alpha}{1-s^2}\sqrt{(1-s^2)^2 + 4s^2} = \frac{\alpha\left(1+s^2\right)}{1-s^2} .$$

2. Using part 1 and Lemma H.1 we get

$$\sqrt{\alpha^2+\frac{\delta^2}{4}}-\frac{\delta}{2}=\frac{\alpha\left(1+s^2\right)}{1-s^2}-\frac{2\alpha s}{1-s^2}=\alpha\frac{\left(1-s\right)^2}{1-s^2}=\alpha\frac{1-s}{1+s}\ .$$

3. Using part 1 and Lemma H.1 we get

$$\sqrt{\alpha^2+\frac{\delta^2}{4}}+\frac{\delta}{2}=\frac{\alpha\left(1+s^2\right)}{1-s^2}+\frac{2\alpha s}{1-s^2}=\alpha\frac{\left(1+s\right)^2}{1-s^2}=\alpha\frac{1+s}{1-s}\ .$$

□

Lemma H.3. *Let*

$$\hat{g}(x)=\frac{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{x}\left(\frac{\delta}{2}+\sqrt{\frac{\delta^2}{4}+x^2}\right)$$

be defined $\forall x>0$, *and* $\forall\delta\geq 0$. *Then:*

$$\lim_{x\to 0^+}\hat{g}(x)=0\ .$$

Proof.

$$\lim_{x\to 0^+}\hat{g}(x)=\lim_{x\to 0^+}\delta\frac{\sqrt{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}}{x}=\lim_{x\to 0^+}\delta\sqrt{\frac{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}{x^2}}\ .$$

Using L'Hopital's rule we have

$$\lim_{x\to 0^+}\frac{\sqrt{x^2+\frac{\delta^2}{4}}-\frac{\delta}{2}}{x^2}=\lim_{x\to 0^+}\frac{\frac{x}{\sqrt{x^2+\frac{\delta^2}{4}}}}{2x}=\lim_{x\to 0^+}\frac{1}{2\sqrt{x^2+\frac{\delta^2}{4}}}=\frac{1}{\delta}\ ,$$

and so

$$\lim_{x\to 0^+}\hat{g}(x)=\lim_{x\to 0^+}\delta\sqrt{\frac{1}{\delta}}=\sqrt{\delta}\ .$$

□

Lemma H.4. *Let* $\mathbf{A}$ *be a positive definite matrix and* $f\left(\mathbf{x}\right)$ *a kernel predictor corresponding to a linear kernel* $K\left(\mathbf{x},\mathbf{x}'\right)=\mathbf{x}^\top\mathbf{A}\mathbf{x}'$. *Then*

$$\|f\|_K^2=\mathbf{w}^\top\mathbf{A}^{-1}\mathbf{w}\ ,$$

where $f\left(\mathbf{x}\right)=\mathbf{w}^\top\mathbf{x}$.

Proof. Write $K\left(\mathbf{x},\mathbf{x}'\right)=\mathbf{x}^\top\mathbf{A}\mathbf{x}'=\mathbf{x}^\top\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{x}'$, then $\phi\left(\mathbf{x}\right)=\mathbf{A}^{\frac{1}{2}}\mathbf{x}$ is the corresponding feature mapping and

$$f\left(\mathbf{x}\right)=\tilde{\mathbf{w}}^\top\phi\left(\mathbf{x}\right)=\tilde{\mathbf{w}}^\top\mathbf{A}^{\frac{1}{2}}\mathbf{x}=\mathbf{w}^\top\mathbf{x}$$

for $\mathbf{w}=\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{w}}$. Therefore

$$\|f\|_K^2=\|\tilde{\mathbf{w}}\|^2=\left\|\mathbf{A}^{-\frac{1}{2}}\mathbf{w}\right\|^2=\mathbf{w}^\top\mathbf{A}^{-1}\mathbf{w}\ .$$

□