# Bound on the Weights of a Neural Network Under Nesterov's Accelerated Gradient Flow

The following is a derivation of a bound on the weights of a vanilla neural network under Nesterov's Accelerated Gradient Flow (the continuous version of Nesterov's Accelerated Gradient Descent). Finding such a bound was stated as an "open question" in Gluch and Urbanke's "Noether" paper (https://arxiv.org/abs/2104.05508). The following derivation assumes familiarity with their paper:

We start with "conservation law" for Nesterov's Accelerated Gradient Flow:

$$\left\langle W^{(h)}, \ddot{W}^{(h)} + \frac{3}{t}\dot{W}^{(h)} \right\rangle - \left\langle W^{(h+1)}, \ddot{W}^{(h+1)} + \frac{3}{t}\dot{W}^{(h+1)} \right\rangle = 0 \tag{1}$$

First multiply by t

$$t\left(\left\langle W^{(h)}, \ddot{W}^{(h)} \right\rangle - \left\langle W^{(h+1)}, \ddot{W}^{(h+1)} \right\rangle\right) + 3\left(\left\langle W^{(h)}, \dot{W}^{(h)} \right\rangle - \left\langle W^{(h+1)}, \dot{W}^{(h+1)} \right\rangle\right) = 0 \tag{2}$$

Notice that the second term can be expressed as a derivative of $||W^{(h)}||_F^2 - ||W^{(h+1)}||_F^2$:

$$t\left(\left\langle W^{(h)}, \ddot{W}^{(h)} \right\rangle - \left\langle W^{(h+1)}, \ddot{W}^{(h+1)} \right\rangle\right) + \frac{3}{2}\frac{\mathrm{d}}{\mathrm{dt}}\left(||W^{(h)}||_F^2 - ||W^{(h+1)}||_F^2\right) = 0 \tag{3}$$

Next, we can perform integration by parts on the first term (showing just layer (h) for simplicity).

$$\int \left\langle W^{(h)}, \ddot{W}^{(h)} \right\rangle t \; dt \tag{4}$$

$$= \int \sum_{i,j} \left(W_{ij}^{(h)} t\right) \ddot{W}_{ij}^{(h)} dt \tag{5}$$

$$= \sum_{i,j} \int \left(W_{ij}^{(h)} t\right) \ddot{W}_{ij}^{(h)} dt \tag{6}$$

$$= \sum_{i,j} \left[-\int \left(\dot{W}_{ij}^{(h)} t + W_{ij}^{(h)}\right) \dot{W}_{ij}^{(h)} dt + W_{ij}^{(h)} \dot{W}_{ij}^{(h)} t\right] \tag{7}$$

$$= \left\langle W^{(h)}, \dot{W}^{(h)} \right\rangle t - \frac{||W^{(h)}||^2}{2} - \int ||\dot{W}^{(h)}||^2 t \; dt \tag{8}$$

Hence, equation (3) becomes

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(||W^{(h)}||_F^2 - ||W^{(h+1)}||_F^2\right) + \frac{\mathrm{d}}{\mathrm{dt}}\left(\left\langle W^{(h)}, \dot{W}^{(h)} \right\rangle t - \left\langle W^{(h+1)}, \dot{W}^{(h+1)} \right\rangle t\right) = t\left(||\dot{W}^{(h)}||^2 - ||\dot{W}^{(h+1)}||^2\right) \tag{9}$$

Now, note that

$$\left\langle W^{(h)}, \dot{W}^{(h)} \right\rangle t = \frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{1}{2}||W^{(h)}||^2 t\right) - \frac{1}{2}||W^{(h)}||^2 \tag{10}$$

Therefore,

$$\frac{\mathrm{d}^2}{\mathrm{dt}^2}\left(t\; ||W^{(h)}||_F^2 - ||W^{(h+1)}||_F^2\right) + \frac{\mathrm{d}}{\mathrm{dt}}\left(||W^{(h)}||_F^2 - ||W^{(h+1)}||_F^2\right) = 2\; t\left(||\dot{W}^{(h)}||^2 - ||\dot{W}^{(h+1)}||^2\right) \tag{11}$$

For simplicity, let $\alpha = ||W^{(h)}||_F^2 - ||W^{(h+1)}||_F^2$. Then,

$$\frac{\mathrm{d}^2}{\mathrm{dt}^2}(t\alpha) + \dot{\alpha} = 2\; t\left(||\dot{W}^{(h)}||^2 - ||\dot{W}^{(h+1)}||^2\right) \tag{12}$$

$$\ddot{\alpha} + \frac{3}{t}\dot{\alpha} = 2\left(||\dot{W}^{(h)}||^2 - ||\dot{W}^{(h+1)}||^2\right) \tag{13}$$

Next, acknowledge that

$$\ddot{\alpha} + \frac{3}{t}\dot{\alpha} \leq 2\left(||\dot{W}^{(h)}||^2 + ||\dot{W}^{(h+1)}||^2\right) \leq 4\left(\frac{1}{2}\left(||\dot{W}^{(h)}||^2 + ||\dot{W}^{(h+1)}||^2\right) + L(\omega) - L(\omega^*)\right) \tag{14}$$

When summed across all layers, the quantity on the right is simply the Hamiltonian of the system. Moreover, the Hamiltonian is decreasing,

$$\dot{\mathcal{H}} = -\frac{3}{t}||\dot{\omega}||^2 \tag{15}$$

It follows, then, that $\mathcal{H} \leq \mathcal{H}_\circ$. Hence,

$$\ddot{\alpha} + \frac{3}{t}\dot{\alpha} \leq 4\mathcal{H}_\circ \tag{16}$$

$$\ddot{\alpha} \leq -\frac{3}{t}\dot{\alpha} + 4\mathcal{H}_\circ \tag{17}$$

Here, we can make the change of variables:

$$\beta = \dot{\alpha} - \mathcal{H}_\circ t \tag{18}$$

$$\dot{\beta} = \ddot{\alpha} - \mathcal{H}_\circ \tag{19}$$

This gives

$$\dot{\beta} \leq -\frac{3}{t}\beta \tag{20}$$

We can now apply Gronwall's Inequality, which yields

$$\beta \leq \beta(t_\circ) e^{\int_{t_\circ}^{t}(-\frac{3}{t})dt} \tag{21}$$

$$\beta \leq \beta(t_\circ)\left(\frac{t_\circ}{t}\right)^3 \tag{22}$$

If we set $t_\circ = 0$, then

$$\dot{\alpha} - \mathcal{H}_\circ t \leq 0 \tag{23}$$

$$\alpha(t) - \alpha(0) \leq \frac{1}{2}\mathcal{H}_\circ t^2 \tag{24}$$

$$\left[\sum_{h=1}^{K-1}\left|||W^{(h)}(t)||_F^2 - ||W^{(h+1)}(t)||_F^2\right|\right] - \left[\sum_{h=1}^{K-1}\left|||W^{(h)}(0)||_F^2 - ||W^{(h+1)}(0)||_F^2\right|\right] \leq \frac{1}{2}\mathcal{H}_\circ t^2 \tag{25}$$

Or, if we note that $\mathcal{H}_\circ = L(\omega(0)) - L^*$ when the velocities are initialized at 0, we arrive at

$$\left[\sum_{h=1}^{K-1}\left|||W^{(h)}(t)||_F^2 - ||W^{(h+1)}(t)||_F^2\right|\right] - \left[\sum_{h=1}^{K-1}\left|||W^{(h)}(0)||_F^2 - ||W^{(h+1)}(0)||_F^2\right|\right] \leq \frac{1}{2}(L(\omega(0)) - L^*)t^2 \tag{26}$$

which looks a lot like the bound that was derived for the case of Newtonian dynamics. The absolute value comes in by observing the symmetry of the problem (e.g. by defining $\alpha$ by its negative and thus achieving a lower bound on the original $\alpha$).