1 Model

We consider a very simple case of the model used in Alessandro's paper. We set $L = 40$ (number of inputs units) and $K = 2$ (number of hidden units). Our analysis is motivated by a model using ReLU activation. However, we will consider a gated deep linear net (GDLN) implementation of the model.

Our model is defined as follows:

$$\hat{y}(x) = \frac{1}{2} \left(g_1(x) W_1 + g_2(x) W_2 \right) x, \tag{1}$$

where g_1 and g_2 are node gates.

We will consider node gates of the form:

$$g_1(x) = \mathbb{1}(\langle x, e_i \rangle \geq 0) \tag{2}$$

$$g_2(x) = \mathbb{1}(\langle x, -e_i \rangle \geq 0). \tag{3}$$

So, $g_1(x) = 1 - g_2(x)$. We call this the "small bump" gate, as the gate is turned on (or off) when the input is positive in the i-th entry. Note i is fixed.

2 Gradient Flow

Recalling the result from the GDLN paper:

$$\tau \frac{d}{dt} W_1 = \frac{1}{2} \left[\Sigma^{yx}(p_1) - W_1 \Sigma^{xx}(p_1, p_1) - W_2 \Sigma^{xx}(p_1, p_2) \right], \tag{4}$$

where

$$\Sigma^{yx}(p) = \left\langle g_p y x^\top \right\rangle_{x,y} \tag{5}$$

$$\Sigma^{xx}(p, q) = \left\langle g_p g_q x x^\top \right\rangle_{x,y}. \tag{6}$$

Note that $\Sigma^{xx}(p_1, p_2) = 0$ by the construction of our gates, since they are never both nonzero. So, we want to compute $\Sigma^{yx}(p_1)$ and $\Sigma^{xx}(p_1, p_1)$.

We compute the former to be:

$$\Sigma^{yx}(p_1) = \frac{1}{\pi} \left[\tan^{-1} \left(\sqrt{\frac{\rho_{ik}^2}{\frac{1}{2g^2} + (1 - \rho_{ik}^2)}} \right) \right]_k^\top, \qquad \rho_{ik} = \exp\left(-\frac{(i-k)^2}{\xi_1^2} \right). \tag{7}$$

The latter is:

$$\Sigma^{xx}(p_1, p_1) = \frac{1}{\pi}\tan^{-1}\left(\sqrt{2}\frac{\rho_{ik} + a_i}{\sqrt{1 + 2a_k^2\sigma_1^2}}\right) \tag{8}$$

$$a_i = \frac{\rho_{kl} - \rho_{il}\rho_{ik}}{1 - \rho_{ik}^2}c \tag{9}$$

$$a_k = \frac{\rho_{il} - \rho_{kl}\rho_{ik}}{1 - \rho_{ik}^2}c \tag{10}$$

$$c = \frac{g}{\sqrt{1 + 2g^2\sigma^2}} \tag{11}$$

$$\sigma^2 = 1 - \frac{1}{1 - \rho_{ik}^2}\left(\rho_{il}^2 - 2\rho_{ik}\rho_{il}\rho_{kl} + \rho_{kl}^2\right) \tag{12}$$

$$= \frac{1}{1 - \rho_{ik}^2}\left(1 - \rho_{ik}^2 - \rho_{il}^2 - \rho_{kl}^2 + 2\rho_{ik}\rho_{il}\rho_{kl}\right) \tag{13}$$

(This looks wrong tbh. See Desmos. First one is right tho.)

(Also I have no idea how to decouple the ODEs with these matrices.)