$L := L(z(t), t, \theta)$

$\dot{z}^\alpha(t) =: f^\alpha(z(t), t, \theta)$

$\Rightarrow z^\alpha(t + \epsilon) = z^\alpha(t) + \epsilon f^\alpha(z(t), t, \theta)$

$\Rightarrow \frac{\partial z^\beta(t+\epsilon)}{\partial z^\alpha(t)} = \delta_\alpha^\beta + \epsilon \frac{\partial f^\beta}{\partial z^\alpha(t)}(z(t), t, \theta)$

$a_\alpha(t) := \frac{\partial L}{\partial z^\alpha(t)}$

$\frac{\partial L}{\partial z^\alpha(t)} = \frac{\partial L}{\partial z^\beta(t+\epsilon)} \frac{\partial z^\beta(t+\epsilon)}{\partial z^\alpha(t)}$

$\Rightarrow a_\alpha(t) = a_\beta(t+\epsilon) \frac{\partial z^\beta(t+\epsilon)}{\partial z^\alpha(t)} = [a_\beta(t) + \epsilon \dot{a}_\beta(t)] \left[\delta_\alpha^\beta + \epsilon \frac{\partial f^\beta}{\partial z^\alpha(t)}(z(t), t, \theta) \right]$

$\Rightarrow \dot{a}_\alpha(t) = -a_\beta(t) \frac{\partial f^\beta}{\partial z^\alpha(t)}(z(t), t, \theta)$

1 Calculation of $\partial L / \partial \theta$

Suppose the model is layerized, the loss depends on the variables (inputs and model parameters) on the ith layer can be regarded as the loss of a new model by truncating the original at the ith layer, which we call $L_i(z_i)$. Varing θ will vary the $L_0(z_0)$ from two aspects, the effect from $\mathrm{dL}_1 / d\theta$ and the Δz_1 caused by $\Delta\theta$.

$\frac{\mathrm{dL}_0}{d\theta}(z_0) = \frac{\mathrm{dL}_1}{d\theta}(z_1) + \frac{\mathrm{dL}_1}{\mathrm{dz}_1} \frac{\partial z_1}{\partial \theta}.$

The same relation holds for any i, by simply considering a truncated model,

$\frac{\mathrm{dL}_i}{d\theta}(z_0) = \frac{\mathrm{dL}_{i+1}}{d\theta}(z_{i+1}) + \frac{\mathrm{dL}_{i+1}}{\mathrm{dz}_{i+1}} \frac{\partial z_{i+1}}{\partial \theta}.$

Thus we have, recursely,

$$\begin{aligned}
\frac{\mathrm{dL}_0}{d\theta}(z_0) &= \frac{\mathrm{dL}_1}{d\theta}(z_1) + \frac{\mathrm{dL}_1}{\mathrm{dz}_1} \frac{\partial z_1}{\partial \theta} \\
&= \left[\frac{\mathrm{dL}_2}{d\theta}(z_1) + \frac{\mathrm{dL}_2}{\mathrm{dz}_2} \frac{\partial z_2}{\partial \theta} \right] + \frac{\mathrm{dL}_1}{\mathrm{dz}_1} \frac{\partial z_1}{\partial \theta} \\
&= \frac{\mathrm{dL}_2}{d\theta}(z_1) + \frac{\mathrm{dL}_2}{\mathrm{dz}_2} \frac{\partial z_2}{\partial \theta} + \frac{\mathrm{dL}_1}{\mathrm{dz}_1} \frac{\partial z_1}{\partial \theta} \\
&= \ldots. \\
&= \frac{\mathrm{dL}_N}{d\theta}(z_1) + \sum_{i=1}^N \frac{\mathrm{dL}_i}{\mathrm{dz}_i} \frac{\partial z_i}{\partial \theta}.
\end{aligned}$$

By $z_{i+1}(t) = z_i(t) + \epsilon f(z_i(t), t; \theta)$, $\partial z_i / \partial \theta = \epsilon \partial f(z_i, t; \theta) / \partial \theta$

1 Continuum of Hopfield

1.1 Hopfield Network

Consider Hopfield network. Let $x(t) \in \{-1, +1\}^N$ denotes the state of the network at descrete time $t = 0, 1, \ldots$; and W a matrix on $\mathbb{R}^N$, essentially ensuring $W_{\alpha\beta} = W_{\beta\alpha}$ and $W_{\alpha\alpha} = 0$. Define energy $E_W(x(t)) := -W_{\alpha\beta} x^\alpha(t) x^\beta(t)$.

Theorem 1. *Along dynamics* $x_\alpha(t+1) = \operatorname{sign}[W_{\alpha\beta} x^\beta(t)]$, $E_W(x(t+1)) - E_W(x(t)) \leqslant 0$.

Proof. Consider the async-updation of Hopfield network. Let's change the component at dimension $\hat{\alpha}$, i.e. $x'_{\hat{\alpha}} = \operatorname{sign}[W_{\hat{\alpha}\beta} x^\beta]$, then

$$\begin{aligned}
E_W(x') - E_W(x) &= -W_{\alpha\beta} x'^\alpha x'^\beta + W_{\alpha\beta} x^\alpha x^\beta \\
&= -2\, (x'^{\hat{\alpha}} - x^{\hat{\alpha}})\, W_{\hat{\alpha}\beta} x^\beta,
\end{aligned}$$

which employs conditions $W_{\alpha\beta}=W_{\beta\alpha}$ and $W_{\alpha\alpha}=0$. Next, we prove that, combining with $x'_{\hat{\alpha}}=\operatorname{sign}[W_{\hat{\alpha}\beta}x^{\beta}]$, this implies $E_W(x')-E_W(x)\leqslant 0$.

If $(x'^{\hat{\alpha}}-x^{\hat{\alpha}})>0$, then $x'^{\hat{\alpha}}=1$ and $x^{\hat{\alpha}}=-1$. Since $x'_{\hat{\alpha}}=\operatorname{sign}[W_{\hat{\alpha}\beta}x^{\beta}]$, $W_{\hat{\alpha}\beta}x^{\beta}>0$. Then $E_W(x')-E_W(x)<0$. Contrarily, if $(x'^{\hat{\alpha}}-x^{\hat{\alpha}})<0$, then $x'^{\hat{\alpha}}=-1$ and $x^{\hat{\alpha}}=1$, implying $W_{\hat{\alpha}\beta}x^{\beta}<0$. Also $E_W(x')-E_W(x)<0$. Otherwise, $E_W(x')-E_W(x)=0$. So, we conclude $E_W(x')-E_W(x)\leqslant 0$. □

Since the states of the network are finite, the E_W is lower bounded. Thus the network converges (relaxes) at finite t.

1.2 Continuum

Let's consider applying the convergence of Hopfield network to neural ODE for generic network architecture. This makes the descrete time t a continuum.

Theorem 2. *Let M be a Riemann manifold with metric g. Given any function $\mathcal{E}\in C^1(M,\mathbb{R})$. Let $x(t)\in C^1(\mathbb{R},M)$ denote trajectory. Then, $d\mathcal{E}/\mathrm{dt}\leqslant 0$ along $x(t)$ if*

$$\frac{dx^{\alpha}}{dt}(t)=-g^{\alpha\beta}(x(t))\frac{\partial\mathcal{E}}{\partial x^{\beta}}(x(t)).$$

Proof. We have

$$\frac{d\mathcal{E}}{dt}(t)=\frac{\partial\mathcal{E}}{\partial x^{\alpha}}(x(t))\frac{dx^{\alpha}}{dt}(t)=-g^{\alpha\beta}(x(t))\frac{\partial\mathcal{E}}{\partial x^{\alpha}}(x(t))\frac{\partial\mathcal{E}}{\partial x^{\beta}}(x(t))\leqslant 0.$$ □

Further, if the function $\mathcal{E}$ is lower bounded, then the trajectory converges (relaxes) at finite t. We call this dynamic with lower bounded $\mathcal{E}$ as "Hopfield dynamic with energy $\mathcal{E}$".

This is the continuum analogy to the convergence of Hopfield network. Indeed, let M be $\mathbb{R}^N$, and $\mathcal{E}(x)=-W_{\alpha\beta}x^{\alpha}x^{\beta}$ with $W_{\alpha\beta}=W_{\beta\alpha}$, then dynamics becomes

$$\frac{dx_{\alpha}}{dt}(t)=-\frac{\partial\mathcal{E}}{\partial x^{\alpha}}(x(t))=-2W_{\alpha\beta}x^{\beta}(t),$$

which makes

$$\frac{d\mathcal{E}}{dt}(t)=-2\frac{dx^{\alpha}}{dt}(t)\,W_{\alpha\beta}x^{\beta}(t).$$

Comparing with the proof of convergence of Hopfield network, i.e. $\Delta E_W(x)=-2\Delta x^{\alpha}W_{\alpha\beta}x^{\beta}$, the analogy is obvious. The only differences are that the condition $W_{\alpha\alpha}=0$ and the sign-function are absent here.

It is known that a Riemann manifold can be locally re-coordinated to be Euclidean. Thus, locally $\exists\hat{x}$ coordinate, s.t.

$$\frac{dx^{\alpha}}{dt}(t)=-\delta^{\alpha\beta}\frac{\partial\mathcal{E}}{\partial x^{\beta}}(x(t)).$$

1.3 Quadratic Form and Homogeneousness

Theorem 3. *Let $h(.;\theta)\in C^1(M,M)$ is a neural network parameterized by θ. If the activations within $h(.;\theta)$ are ReLU and linear ones, then the Hopfield dynamic with a quadratic form of energy $\mathcal{E}(x)=h_{\alpha}(x;\theta)h^{\alpha}(x;\theta)$ is formally invariant for rescaling $x\to\lambda x$.*

Proof. If $\mathcal{E}(x) = h_\alpha(x;\theta) h^\alpha(x;\theta)$, then the Hopfield dynamics with energy $\mathcal{E}$ becomes

$$\frac{dx_\alpha}{dt}(t) = -2 h_\beta(x(t);\theta) \nabla_\alpha h^\beta(x(t);\theta).$$

If $h(.;\theta)$ is homogeneous, i.e. $\exists r \in \mathbb{R}\ \forall \lambda \in \mathbb{R}\ h(\lambda x;\theta) = \lambda^r h(x;\theta)$, then $x \to \lambda x$ induces

$$\frac{dx_\alpha}{dt}(t) = -2\lambda^{2r-2} h_\beta(x(t);\theta) \nabla_\alpha h^\beta(x(t);\theta).$$

If $r = 1$, then this dynamic is formally invariant for rescaling $x \to \lambda x$.

Notice that ReLU activation is homogeneous with $r = 1$, so is any linear transform. Then general neural network with ReLU and linear activations is homogeneous with $r = 1$. For such neural networks, the dynamic is formally invariant. $\square$