1 Adjoint Method

Let M a manifold, and $x(t) \in C^1(\mathbb{R}, M)$ a trajectory, obeying

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = f(t, x(t); \theta),$$

and

$$x(t_0) = x_0,$$

where $f \in C(\mathbb{R} \times M, T_M)$ parameterized by θ. For $\forall t_1 > t_0$, let

$$x_1 := x_0 + \int_{t_0}^{t_1} f(t, x(t); \theta) \mathrm{d}t.$$

Then

Theorem 1. *Let $\mathcal{C} \in C^1(M, \mathbb{R})$, and $\forall x(t) \in C^1(\mathbb{R}, M)$ obeying dynamics $f(t, x; \theta) \in C^1(\mathbb{R} \times M, T_M)$ with initial value $x(t_0) = x_0$. Denote*

$$L := \mathcal{C}\left(x_0 + \int_{t_0}^{t_1} f(\tau, x(\tau); \theta) \mathrm{d}\tau \right).$$

Then we have, for $\forall t \in [t_0, t_1]$ given,

$$\frac{\partial L}{\partial x(t)} = \frac{\partial L}{\partial x_1} - \int_t^{t_1} \frac{\partial L}{\partial x^\beta(\tau)} \frac{\partial f^\beta}{\partial x^\alpha(\tau)}(\tau, x(\tau); \theta) \mathrm{d}\tau,$$

and

$$\frac{\partial L}{\partial \theta} = -\int_{t_0}^{t_1} \frac{\partial L}{\partial x^\beta(\tau)} \frac{\partial f^\beta}{\partial \theta}(\tau, x(\tau); \theta) \mathrm{d}\tau.$$

Proof. Suppose the $x(t)$ is layerized, the L 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 $L_i(z_i)$.

$$\begin{aligned}
\frac{\partial L_i}{\partial x_i^\alpha}(x_i) &= \frac{\partial L_{i+1}}{\partial x_{i+1}^\beta}(x_{i+1}) \frac{\partial x_{i+1}^\beta}{\partial x_i^\alpha}(x_i) \\
&= \frac{\partial L_{i+1}}{\partial x_1^\beta}(x_{i+1}) \frac{\partial}{\partial x_i^\alpha}(x_i^\beta + f^\beta(t_i, x_i; \theta)\Delta t) \\
&= \frac{\partial L_{i+1}}{\partial x_{i+1}^\alpha}(x_{i+1}) + \frac{\partial L_{i+1}}{\partial x_{i+1}^\beta}(x_{i+1}) \partial_\alpha f^\beta(t_i, x_i; \theta)\Delta t.
\end{aligned}$$

This hints that

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial x^\alpha(t)} = -\frac{\partial L}{\partial x^\beta(t)} \frac{\partial f^\beta}{\partial x^\alpha(t)}(t, x(t); \theta).$$

The initial value is $\partial L / \partial x_1$. Thus

$$\frac{\partial L}{\partial x(t)} = \frac{\partial L}{\partial x_1} - \int_t^{t_1} \frac{\partial L}{\partial x^\beta(\tau)} \frac{\partial f^\beta}{\partial x^\alpha(\tau)}(\tau, x(\tau); \theta) \mathrm{d}\tau.$$

Varing θ will vary the $L_i(x_i)$ from two aspects, the effect from $\partial L_{i+1}/\partial\theta$ and the Δx_{i+1} caused by $\Delta\theta$.

$$\begin{aligned}\frac{\partial L_i}{\partial\theta}(x_i) =& \frac{\partial L_{i+1}}{\partial\theta}(x_{i+1}) + \frac{\partial L_{i+1}}{\partial x_{i+1}}\frac{\partial x_{i+1}}{\partial\theta} \\ =& \frac{\partial L_{i+1}}{\partial\theta}(x_{i+1}) + \frac{\partial L_{i+1}}{\partial x_{i+1}}\frac{\partial}{\partial\theta}(x_i^\beta + f^\beta(t_i, x_i;\theta)\Delta t) \\ =& \frac{\partial L_{i+1}}{\partial\theta}(x_{i+1}) + \frac{\partial L_{i+1}}{\partial x_{i+1}}\frac{\partial f^\beta}{\partial\theta}(t_i, x_i;\theta)\Delta t.\end{aligned}$$

This hints that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\theta} = -\frac{\partial L}{\partial x^\alpha(t)}\frac{\partial f^\beta}{\partial\theta}(t, x(t), \theta).$$

The initial value is 0 since $\mathcal{C}(.)$ is explicitly independent on θ. Thus

$$\frac{\partial L}{\partial\theta} = -\int_{t_0}^{t}\frac{\partial L}{\partial x^\beta(\tau)}\frac{\partial f^\beta}{\partial\theta}(\tau, x(\tau);\theta)\mathrm{d}\tau. \qquad \square$$

2 Continuous-time Hopfield Network

2.1 Discrete-time 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 2. *Along dynamics* $x_\alpha(t+1) = \mathrm{sign}[W_{\alpha\beta}\, x^\beta(t)]$, $E_W(x(t+1)) - E_W(x(t)) \leqslant 0$.

Proof. Consider the async-updation of Hopfield network. Let's change the component at dimension $\hat{\alpha}$, i.e. $x'_{\hat{\alpha}} = \mathrm{sign}[W_{\hat{\alpha}\beta}\, x^\beta]$, 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 $x'_{\hat{\alpha}} = \mathrm{sign}[W_{\hat{\alpha}\beta}\, x^\beta]$, 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 $x'_{\hat{\alpha}} = \mathrm{sign}[W_{\hat{\alpha}\beta}\, x^\beta]$, $W_{\hat{\alpha}\beta}\, x^\beta > 0$. 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 $E_W(x') - E_W(x) \leqslant 0$. $\square$

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

2.2 Continuum of Time

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

Theorem 3. *Let M be a Riemann manifold. Given $\mathcal{E} \in C^1(M, \mathbb{R})$. For $\forall x(t) \in C^1(\mathbb{R}, M)$ s.t.*

$$\frac{dx^\alpha}{dt}(t) = -\nabla^\alpha \mathcal{E}(x(t)),$$

then $d\mathcal{E}/\mathrm{dt} \leqslant 0$ along $x(t)$. Further, if $\mathcal{E}$ is lower bounded, then $\exists t_\star < +\infty$, s.t. $\mathrm{d}x^\alpha/\mathrm{dt} = 0$ at $t_\star$.

Proof. We have

$$\frac{d\mathcal{E}}{dt}(t) = \nabla_\alpha \mathcal{E}(x(t)) \frac{dx^\alpha}{dt}(t) = -\nabla_\alpha \mathcal{E}(x(t)) \nabla^\alpha \mathcal{E}(x(t)) \leqslant 0. \qquad \square$$

Remark 4. 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

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

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.

2.3 Energy as Neural Network

The funciton $\mathcal{E}$ is the energy in the Ising model (as a toy Hopfield network).

Theorem 5. *Let f_θ a neural network mapping from M to $\mathbb{R}$, parameterized by θ, and $\mathcal{B}: M \to D$ where $D \subseteq M$ being compact. Then*

$$\mathcal{E}_\theta := f_\theta \circ \mathcal{B}$$

is a bounded function in $C^1(M, \mathbb{R})$.

One option of G is tanh-function. However, the $\tanh(x)$ will be saterated as $x \to \pm\infty$. A better option is *boundary reflection*. Define boundary reflection map

$$f_{\mathrm{BR}}: \mathbb{R}^d \to [0, 1]^d$$
$$f_{\mathrm{BR}}(x) = \begin{cases} x, x \in [0, 1] \\ -x, x \in [-1, 0] \\ f_{\mathrm{BR}}(x-2), x > 1 \\ f_{\mathrm{BR}}(x+2), x < -1 \end{cases}.$$

This function has constant gradient ± 1, thus no saturation. It has periodic symmetry.