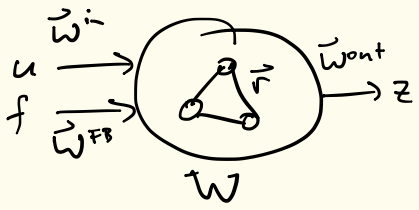
2019.05.15

Stability in trained RNNs

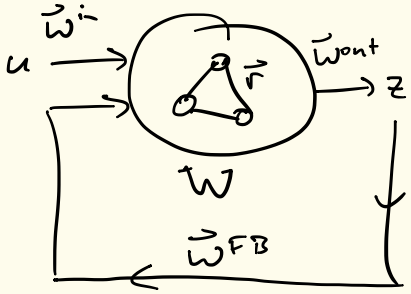
Ref: Rivkind and Barak (PRL, 2017)

Train RNN with the echo state approach similar to Jaeger + Haas (2004).

Train in open loop:



Test in closed loop:



$$\dot{\vec{x}} = -\vec{x} + W \cdot \vec{r} + \vec{w}^{FB} f + \vec{w}^{in} u \qquad (1)$$

$$\dot{\vec{x}} = -\vec{x} + W \cdot \vec{r} + \vec{w}^{FB} z + \vec{w}^{in} u \qquad (2)$$

where $\vec{r} = \phi(\vec{x})$, $z = \vec{w}^{out} \cdot \vec{r}$, and $W_{ij} \sim \mathcal{N}(0, \frac{g^2}{N})$.

Training is done with least squares so $z \approx f$.

Note: This differs from FORCE learning, where RLS is used rather than LS, and the training is done in closed loop.

Main question: If we train the RNN to produce a fixed point, $f(t) = A$, will it be stable when we close the loop?

(2)

Assume $u=0$. Then the open-loop fixed point from (1) is, for $f(t)=A$,

$$\vec{\bar{x}} = W \cdot \phi(\vec{\bar{x}}) + \vec{w}^{FB} A \qquad (3)$$

Is this solution stable? Let $\vec{x} = \vec{\bar{x}} + \delta\vec{x}(t)$ in (1) and linearize:

$$\begin{aligned}
\delta\dot{x}_i &= -(\bar{x}_i + \delta x_i) + \sum_j W_{ij}\,\phi(\bar{x}_j + \delta x_j) + w^{FB}_i A \\
&= -(\bar{x}_i + \delta x_i) + \sum_j W_{ij}\,\phi(\bar{x}_j) + \sum_j W_{ij}\,\phi'(\bar{x}_j)\,\delta x_j \\
&\quad + w^{FB}_i A \\
&\overset{(3)}{=} -\delta x_i + \sum_j W_{ij}\,\phi'(\bar{x}_j)\,\delta x_j \\
&= \sum_j P_{ij}\,\delta x_j, \qquad P_{ij} = -\delta_{ij} + W_{ij}\,\phi'(\bar{x}_j) \qquad (4)
\end{aligned}$$

The solution is stable if all evals of P have real parts ≤ 0. From random matrix theory, this happens if $g^2 \langle (\phi'(\bar{x}_j))^2 \rangle \le 1$, where $\langle \cdots \rangle$ is over neurons.

Is the solution (3) stable if we close the loop? Then rather than (4) we have

$$\delta\dot{x}_i = \sum_j \overbrace{\left[-\delta_{ij} + \left(W_{ij} + w^{FB}_i w^{out}_j\right)\phi'(\bar{x}_j)\right]}^{Q_{ij}} \delta x_j \qquad (5)$$

Now we need to compute evals of Q. To do this, first return to the open-loop case and compute the response $\delta\vec{x}(t)$ to a perturbation $\delta f(t) = f(t) - A$ (cf. Eq. (4)):

$$\delta\dot{x}_i = -\delta x_i + \sum_j W_{ij}\,\phi'(\bar{x}_j)\,\delta x_j + w_i^{FB}\,\delta f \qquad (6)$$

Let

$$\delta\vec{x}(t) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{i\omega t} \underline{X}_i(\omega)$$

$$\delta f(t) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{i\omega t} F(\omega) \qquad (7)$$

$$\delta z(t) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{i\omega t} Z(\omega)$$

(6)

$$\Rightarrow i\omega \underline{X}_i(\omega) = -\underline{X}_i(\omega) + \sum_j W_{ij}\,\phi'(\bar{x}_j)\,\underline{X}_j(\omega) + w_i^{FB} F(\omega) \qquad (8)$$

The perturbation of the output is

$$\delta z(t) = \sum_i w_i^{out}\,\phi'(\bar{x}_i)\,\delta x_i(t) \qquad (9)$$

$$\Rightarrow Z(\omega) = \sum_i w_i^{out}\,\phi'(\bar{x}_i)\,\underline{X}_i(\omega) \qquad (10)$$

(4) The open-loop gain:

$$G^{OL}(\omega) = \frac{\bar{z}(\omega)}{F(\omega)}$$

$$= \frac{1}{F(\omega)} \sum_i \omega_i^{out} \phi'(\bar{x}_i) \bar{X}_i(\omega) \tag{11}$$

The closed-loop gain is then

$$G^{CL}(\omega) = \frac{\bar{z}}{F - \bar{z}} = \frac{G^{OL}(\omega)}{1 - G^{OL}(\omega)} \tag{12}$$

We want to find the poles of $G^{CL}(\omega)$. First we need $G^{OL}(\omega)$.

In (3), let

$$\bar{x}_i = \bar{x}_i^0 + \bar{x}_i^1 \tag{13}$$

where

$$\bar{x}_i^0 = \omega_i^{FB} A \tag{14}$$

$$\bar{x}_i^1 = \sum_j W_{ij} \phi(\bar{x}_j) \tag{15}$$

Then, as in DMFT, assume $\bar{x}_i^1 \sim \mathcal{N}(0, \sigma^2)$, where, from (15), σ is given self-consistently:

$$\langle (\bar{x}_i^1)^2 \rangle = \left\langle \sum_{jk} W_{ij} W_{ik} \phi(\bar{x}_j^0 + \bar{x}_j^1) \phi(\bar{x}_k^0 + \bar{x}_k^1) \right\rangle$$

$$= \sum_{jk} \langle W_{ij} W_{ik} \rangle \left\langle \phi(\bar{x}_j^0 + \bar{x}_j^1) \phi(\bar{x}_k^0 + \bar{x}_k^1) \right\rangle$$

$$= g^2 \left\langle \phi^2(\omega_j^{FB} A + \bar{x}_j^1) \right\rangle$$

(5) $\Rightarrow \sigma^2 = g^2 \langle \phi^2(w_j^{FB} A + \sigma y)\rangle$ (16)

$= g^2 \int dw^{FB}\, p(w^{FB}) \int \frac{dy}{\sqrt{2\pi}} e^{-y^2/2} \phi^2(w^{FB} A + \sigma y)$

Similarly, let

$\vec{\bar{x}}(\omega) = \vec{\bar{x}}^0 + \vec{\bar{x}}^1$ (17)

where, from (8),

$\bar{x}^0_i(\omega) = \frac{1}{1+i\omega} w_i^{FB} F(\omega)$ (18)

$\bar{x}^1_i(\omega) = \frac{1}{1+i\omega} \sum_j w_{ij} \phi'(\bar{x}_j) \bar{x}^1_j(\omega)$ (19)

Now compute the overlap of $\vec{\bar{x}}^1$ with $\vec{\bar{x}}^1$:

$(1+i\omega) \sum_i \bar{x}^{1*}_i \bar{x}^1_i \overset{(15,19)}{=} \sum_{ijk} w_{ij} \phi(\bar{x}_j) w_{ik} \phi'(\bar{x}_k) \bar{x}^1_k$

$\overset{(18,19)}{=} \sum_{ijk} w_{ij} w_{ik} \phi(\bar{x}_j) \phi'(\bar{x}_k) \left[w_k^{FB} \frac{F(\omega)}{1+i\omega} + \bar{x}^1_k \right]$ (20)

Further decompose $\vec{\bar{x}}^1 = \vec{\bar{x}}^{\parallel} + \vec{\bar{x}}^{\perp}$, where

$\bar{x}^{\parallel}_i(\omega) = \alpha(\omega) \bar{x}^1_i$ (21)

Then (20) is

$(1+i\omega)\alpha(\omega) |\vec{\bar{x}}^1|^2 = \sum_{ijk} w_{ij} w_{ik} \phi(\bar{x}_j) \phi'(\bar{x}_k)$

$\times \left[w_k^{FB} \frac{F(\omega)}{1+i\omega} + \alpha(\omega) \bar{x}^1_k + \bar{x}^{\perp}_k(\omega) \right]$

⑥ Taking $\langle \cdots \rangle$ of this gives

$$(1+i\omega)\alpha(\omega)N\sigma^2 = g^2 \sum_k \Big\langle \phi(\bar{x}_k)\, \phi'(\bar{x}_k) \times \Big[w_k^{FB} \frac{F(\omega)}{1+i\omega} + \alpha(\omega)\bar{x}_k^{\perp} \Big] \Big\rangle \tag{22}$$

[The last term vanished "because $\vec{\bar{x}}^{\perp}$ is indep. of $\vec{\bar{x}}^{\parallel}$, and thus, on average, perp. to any function of $\vec{\bar{x}}^{\parallel}$."]

Following (13)-(16), replace $\bar{x}^{\parallel}$ by random variables:

$$\bar{x}_i = wA + \sigma y$$

where $w \sim p(w)$ and $y \sim \mathcal{N}(0,1)$.

$$\overset{(22)}{\Rightarrow} (1+i\omega)\alpha(\omega) = \frac{g^2}{\sigma^2} \int Dw \int Dy\, \phi(wA+\sigma y)\, \phi'(wA+\sigma y) \times \Big[w \frac{F(\omega)}{1+i\omega} + \alpha(\omega)\sigma y \Big]$$

$$\equiv \frac{\beta_0 F(\omega)}{1+i\omega} + \beta_1 \alpha(\omega)$$

$$\Rightarrow \alpha(\omega) = \frac{\beta_0 F(\omega)}{(1+i\omega)(1+i\omega-\beta_1)} \tag{23}$$

To get $\vec{w}^{out}$, note that $\vec{w}^{out} \cdot \phi(\vec{\bar{x}}) = A$ is solved by

$$\vec{w}^{out} = \frac{A}{|\phi(\vec{\bar{x}})|^2} \phi(\vec{\bar{x}}) \simeq \frac{A}{N\langle \phi^2(\bar{x}_i) \rangle} \phi(\vec{\bar{x}}) \overset{(16)}{=} \frac{g^2}{N\sigma^2} A \phi(\vec{\bar{x}}) \tag{24}$$

(7) Using this with (11) gives

$$G^{OL}(\omega) = \frac{1}{F(\omega)} \sum_i \langle w_i^{out} \quad \phi'(\bar{x}_i) \underline{X}_i(\omega) \rangle$$

$$= \frac{g^2 A}{N\sigma^2 F(\omega)} \sum_i \langle \phi(\bar{x}_i) \phi'(\bar{x}_i) \underline{X}_i(\omega) \rangle$$

$$= \frac{g^2 A}{\sigma^2 F(\omega)} \langle \phi(\bar{x}_i) \phi'(\bar{x}_i) \underline{X}_i(\omega) \rangle$$

$$\overset{(22)}{=} \frac{A}{N\sigma^2 F(\omega)} N\sigma^2 (1+i\omega)\alpha(\omega)$$

$$= \frac{A}{F(\omega)} \alpha(\omega)(1+i\omega)$$

$$\overset{(23)}{=} \frac{A\beta_0}{1+i\omega-\beta_1} \tag{25}$$

Then, from (12),

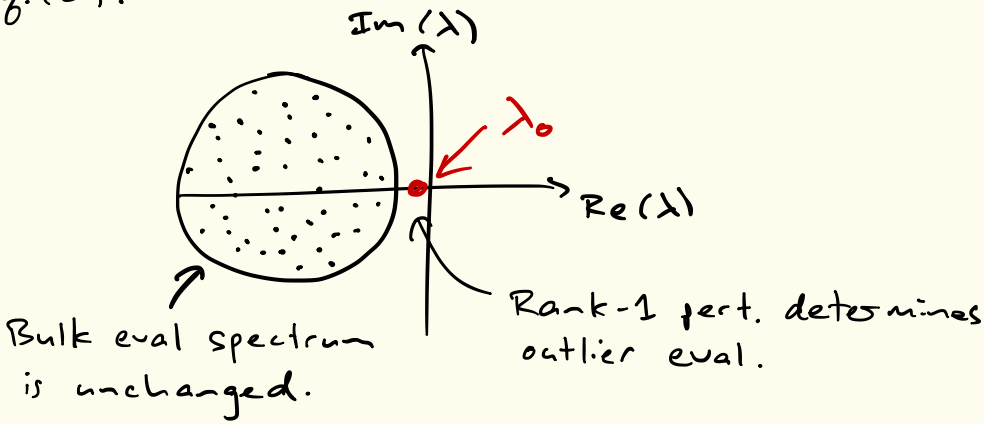
$$G^{CL}(\omega) = \frac{1}{\frac{1+i\omega-\beta_1}{A\beta_0} - 1}$$

$$= \frac{A\beta_0}{i\omega - A\beta_0 - \beta_1 + 1} \tag{26}$$

This expression has a single pole at

$$\lambda_0 = A\beta_0 + \beta_1 - 1 \qquad (27)$$

This is the outlier eigenvalue of the effective CL connectivity matrix Q from Eq. (5).



Recipe:

1) Given g and A, calculate σ self-consistently from (16).
2) With g, A, and σ, calculate β_0 and β_1 (p. 6).
3) Calculate λ_0 from (27).

$\lambda_0 < 0 \Rightarrow$ stable

$\lambda_0 > 0 \Rightarrow$ unstable

Generalization to M outputs is possible. Resonance frequencies (large $G^{cc}(\omega = \omega_0)$) can occur in this case.

Relation to Mastrogiuseppe + Ostojic (2019)

M+O: Given overlaps between $\vec{w}^{in}$, $\vec{w}^{out}$ and $\vec{w}^{FB}$, as well as target A, use DMFT to solve self-consistently for $\vec{w}^{out}$ and $\Delta \equiv \mathrm{Var}(x_i)$. Then

1) Check stability of the $z = A$ fixed point by computing the outlier eigenvalue with DMFT (M+O, 2018).

2) Check for other fixed points and test their stability.

Notes:

i) Since DMFT doesn't require linearization, the fixed-point analysis is <u>global</u>, unlike Rivkind's.

ii) DMFT, unlike Rivkind, assumes the recurrent input is random noise.