Gaussian-process Bayesian inference in a recurrent network

Moritz Helias

February 12, 2020

Abstract

These notes contain a field-theoretical formulation of Lee et al. 2018 and some details of the calculation for the Bayesian inference.

1 Path-integral formulation of network dynamics

Network dynamics of a network of $i = 1, \ldots, N$ neurons

$$h_i(t+t) = \sum_j w_{ij}\phi(x_j(t)) - \tilde{j}_i(t), \tag{1}$$

where $\phi = \tanh$ is the gain function and $-\tilde{j}$ is an input supplied to the network. We write as generating functional for the moments of the input

$$Z[j](w) = \int \mathcal{D}h\, \delta\big[h(\circ+1) - w\phi(h(\circ)) + \tilde{j}(\circ)\big]\ \exp(j^{\mathrm{T}}h),$$

where $\circ$ stands for the time argument of the functions and $[w\phi(h(\circ))]_i = \sum_j w_{ij}\phi(x_j(\circ))$ is to be understood element-wise.

Introduce auxiliary fields for express the Dirac-δ as by its Fourier transform

$$\delta[h] = \int \mathcal{D}\tilde{h}\ \exp\left(\tilde{h}^{\mathrm{T}}h\right)$$

with $\int \mathcal{D}\tilde{h} = \prod_t \int_{-i\infty}^{i\infty} \frac{d\tilde{h}(t)}{2\pi i}$ and $\tilde{h}^{\mathrm{T}}h = \sum_{i,t} \tilde{h}_i(t)h_i(t)$ is the inner product over units and time. Then Z takes the form

$$Z[j](w) = \int \mathcal{D}h \int \mathcal{D}\tilde{h}\ \exp\left(\tilde{h}^{\mathrm{T}}\left[h - w\phi(h)\right)\right] + j^{\mathrm{T}}h + \tilde{j}^{\mathrm{T}}\tilde{h}\big).$$

So the input $\tilde{j}$ formally plays the role

The output of the network will be $h(T)$ at some time T. We are interested in the distribution of this output taken over weights w.

Disorder average over $w \sim \mathcal{N}(0, \frac{g^2}{N})$ yields (using Hubbard-Stratonovich / Gaussian identity $\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}y^2 + xy} = e^{\frac{1}{2}x^2}$)

$$\begin{aligned}
\langle \exp\big(\tilde{h}^{\mathrm{T}} w \phi(h)\big) \rangle_J &= \exp\big(\frac{g^2}{2N} \sum_{i,j,t,t'} \tilde{h}_i(t)\tilde{h}_i(t')\, \phi(h_j(t))\phi(h_j(t'))\big) \\
&= \exp\big(\frac{g^2}{2N} \sum_{t,t'} \sum_i \tilde{h}_i(t)\tilde{h}_i(t') \sum_j \phi(h_j(t))\phi(h_j(t'))\big),
\end{aligned}$$

where the appearance of sums over all units show that after the average the units become all statistically identical.

Introducing auxiliary field

$$Q_1(t,t') := \frac{g^2}{N} \sum_j \phi(h_j(t))\phi(h_j(t'))$$

and helping field Q_2 to express latter constraint, we get

$$\begin{aligned}
\langle Z[j,\tilde{j}](w) \rangle_w = &\int \mathcal{D}Q_1 \int \mathcal{D}Q_2 \, \exp\big(-\frac{N}{g^2} Q_1^{\mathrm{T}} Q_2 \\
&\times \int \mathcal{D}h \int \mathcal{D}\tilde{h} \, \exp\big(\tilde{h}^{\mathrm{T}} h + \frac{1}{2}\tilde{h}^{\mathrm{T}} Q_1 \tilde{h} + \phi(h)^{\mathrm{T}} Q_2 \phi(h) + j^{\mathrm{T}} h + \tilde{j}^{\mathrm{T}} \tilde{h}\big).
\end{aligned}$$

Latter line factorizes in the unit index, so we get the same integral to the power of N

$$\begin{aligned}
\langle Z[j,\tilde{j}](w) \rangle_w &= \int \mathcal{D}Q_1 \int \mathcal{D}Q_2 \, \exp\big(-\frac{N}{g^2} Q_1^{\mathrm{T}} Q_2 + N\,\Omega(Q_1,Q_2,j,\tilde{j})\big), \\
\Omega(Q_1,Q_2,j,\tilde{j}) &:= \ln \int \mathcal{D}h_1 \int \mathcal{D}\tilde{h}_1 \\
&\quad \times \exp\big(\tilde{h}_1^{\mathrm{T}} h_1 + \frac{1}{2}\tilde{h}_1^{\mathrm{T}} Q_1 \tilde{h}_1 + \phi(h_1)^{\mathrm{T}} Q_2 \phi(h_1) + j_1^{\mathrm{T}} h_1 + \tilde{j}_1^{\mathrm{T}} \tilde{h}_1\big),
\end{aligned}$$

where the latter integral is only over a scalar field h_1 and $\tilde{h}_1$.

We perform a saddle-point approximation in Q_1 and Q_2 to obtain the stationarity equations

$$\begin{aligned}
0 &\overset{!}{=} \frac{\delta}{\delta Q_1}\big(-\frac{N}{g^2} Q_1^{\mathrm{T}} Q_2 + N\,\Omega(Q_1,Q_2)\big) \quad \to \quad Q_2^*(t,t') = \frac{1}{2}\langle \tilde{h}_1(t)\tilde{h}_1(t') \rangle \equiv 0, \\
0 &\overset{!}{=} \frac{\delta}{\delta Q_2}\big(-\frac{N}{g^2} Q_1^{\mathrm{T}} Q_2 + N\,\Omega(Q_1,Q_2)\big) \quad \to \quad Q_1^*(t,t') = g^2 \langle \phi(h_1(t))\phi(h_1(t')) \rangle,
\end{aligned}$$

where the latter expectation value is with regard to the action $\Omega(Q_1^*, Q_2^*)$. The saddle point value of Q_2 vanishes by the normalization condition.

Because at the saddle point $\Omega(Q_1^*, 0)$ describes Gaussian statistics of the h. This can be seen by performing the integral over $\tilde{h}$

$$\int \mathcal{D}\tilde{h}_1 \, \exp\big(\tilde{h}_1^{\mathrm{T}} h_1 + \frac{1}{2}\tilde{h}_1^{\mathrm{T}} Q_1 \tilde{h}_1\big) = \exp\big(-\frac{1}{2} h^{\mathrm{T}} Q_1^{-1} h\big),$$

we have that the latter saddle point equation is an double integral over jointly Gaussian distributed $h_1(t)$ and $h_1(t')$

$$Q_1^*(t,t') = g^2 \, \langle \phi(h_1(t)) \, \phi(h_1(t')) \rangle.$$

We need the time-evolution of

$$\begin{aligned} q(t) := Q_1^*(t,t) &= g^2 \, \langle \phi(h_1) \, \phi(h_1) \rangle \\ &= g^2 \, \langle \phi(h) \phi(h) \rangle_{h_1 \sim \mathcal{N}\left(0, q(t-1)\right)}. \end{aligned}$$

The input x is applied in the first time step $t = 0$ by setting

$$\tilde{j}(0) = -x.$$

By (1) this sets the input to the initial values, where we assume that $x(t < 0) = 0$.

We thus need to compute the iteration

$$\text{for } t = 1: \quad q(t+1) = \begin{cases} g^2 \, \phi(x)\phi(x) & \text{for } t = 0 \\ g^2 \, \langle \phi(h)\phi(h) \rangle_{h \sim \mathcal{N}(0, q(t))} & \text{for } t > 0 \end{cases}.$$

The output y of the network if then given by $y = h(T)$ for some time point T. So the mapping of an input x to the output y can be written as

$$\begin{gathered} y = f_w(x), \\ \text{where } y := h(T) \text{ for } h(0) = x. \end{gathered}$$

The distribution of outputs over the realization of the random connections is a Gaussian

$$\int p(y|w,x) \, p(w) \, dw = \mathcal{N}(0, q(T)).$$

$$\begin{aligned} \langle y_i \rangle &= 0 \\ \langle y_i y_j \rangle &= \langle h_i(T) h_j(T) \rangle = \delta_{ij} \, q(T). \end{aligned}$$

Likewise, we define the joint distribution for two different inputs

$$p(y, y'|x, x') = \mathcal{N}(0, K_{xx'}(T)),$$

where $K_{xx'}(t)$ satisfies the recurrence relation

$$K_{xx'}(t+1) = \begin{cases} g^2\phi(x)\phi(x') & \text{for } t=0 \\ g^2\langle\phi(h)\phi(h)\rangle_{h\sim\mathcal{N}(0,K_{xx'}(t))} & \text{for } t>0 \end{cases}. \tag{2}$$

This iteration can be derived from the replica calculation, starting from the moment-generating function

$$\begin{aligned} Z[j^{(1)},j^{(2)}](w) = \int \mathcal{D}h^{(1)} \int \mathcal{D}h^{(2)}\, &\delta\big[h^{(1)}(\circ+1) - w\phi(h^{(1)}(\circ)) + \tilde{j}^{(1)}(\circ)\big] \\ \times\, &\delta\big[h^{(2)}(\circ+1) - w\phi(h^{(2)}(\circ)) + \tilde{j}^{(2)}(\circ)\big] \\ &\exp(j^{(1)\mathrm{T}}h^{(1)} + j^{(2)\mathrm{T}}h^{(2)}), \end{aligned}$$

which describes a pair of systems with identical connectivity w but different inputs $\tilde{j}^{(1)}$ and $\tilde{j}^{(2)}$. Similar steps as outlined above (disorder average, saddle point approximation) then yield (2).

1.1 Bayesian prediction for output

Assume we have an input-output mapping $y = f_w(x)$, where w are the weights to be trained, x is the input to the network and y its output. We have training data (x_D, y_D) consisting of inputs x_D and outputs y_D. The prior distribution of the weights is $p(w)$. This represents some knowledge we may have about the physical range of these parameters; for example they should not be arbitrarily large. Then Bayes theorem states

$$\begin{aligned} p(y|w)\,p(w) &= p(w|y)\,p(y) \\ p(w|y) &= p(y|w)\,\frac{p(w)}{p(y)}. \end{aligned} \tag{3}$$

The latter is the distribution of the weights that we have after having presented the training data y.

The distribution of outputs given the weights is

$$p(y|w) = \delta(y - f_w(x)).$$

The distribution of the output y_* for a new input x_* is

$$\begin{aligned} p(y_*|y) &= \int p(w|y)\,\delta(y_* - f_w(x_*))\,dw \\ &= \frac{1}{p(y)} \int \delta(y - f_w(x))\,\delta(y_* - f_w(x_*))\,p(w)\,dw. \end{aligned} \tag{4}$$

But the latter integral is just the joint distribution of the output for the training data and the test point

$$p(y_*, y|x^*, x) = \int \delta(y_* - f_w(x_*))\,\delta(y - f_w(x))\,p(w)\,dw.$$

If this distribution is Gaussian, namely if

$$p(y, y^*) = \mathcal{N}(0, K)$$
$$K = \begin{pmatrix} K_{DD} & K_{D*} \\ K_{*D} & K_{**} \end{pmatrix}$$

the distribution (4) is Gaussian, too, namely

$$p(y, y^*) = N \, \exp\Big(-\frac{1}{2}(y, y_*)^{\mathrm{T}} K^{-1} (y, y^*)\Big).$$

The inverse matrix K^{-1} is (see e.g. https://en.wikipedia.org/wiki/Block_matrix; here we mixed both representations that are given there)

$$K^{-1} = \begin{pmatrix} (K_{DD} - K_{D*}K_{**}^{-1}K_{*D})^{-1} & -K_{DD}^{-1}K_{D*}(K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1} \\ -(K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1}K_{*D}K_{DD}^{-1} & (K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1} \end{pmatrix}.$$

By inserting y, the y^*-dependent part of the density is hence also Gaussian, namely

$$\propto \exp\Big(-\frac{1}{2}y^{*\mathrm{T}}(K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1}y^*$$
$$+\frac{1}{2}y^{*\mathrm{T}}\,(K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1}K_{*D}K_{DD}^{-1}\,y + \frac{1}{2}y^{\mathrm{T}}\,K_{DD}^{-1}K_{D*}(K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1}\,y^*\Big)$$

$$= \exp\Big(-\frac{1}{2}(y^{*\mathrm{T}} - K_{*D}K_{DD}^{-1}y)\,(K_{**} - K_{*D}K_{DD}^{-1}K_{D*})^{-1}(y^* - K_{*D}K_{DD}^{-1}y)\Big)\,f(y).$$

So we read off the mean and variance from the last line as (similar as in Lee 2017, eq. (8), (9); they have some readout noise in addition)

$$\mu_{y^*} = K_{*D}K_{DD}^{-1}y,$$
$$\sigma_{y^*}^2 = (K^{-1})_{**}^{-1} = K_{**} - K_{*D}K_{DD}^{-1}K_{D*}.$$

These expressions yield the mean output $\langle y_* \rangle = \mu_{y_*}$ and its uncertainty $\sigma_{y_*}^2$, both for input x_*.

1.2 Questions

- Plan: start with simple example of Bayesian inference e.g. applied to linear regression $y = m\,x$.
- It is known that the network transitions to chaos at the point where $g^2 = 1$. Is this point particularly good for computation?
- How does performance depend on the time T (depth of signal propagation) within which the network processes the input?
- How do the networks perform if the units are noisy?
- What is the distribution of weights after training? It can be computed from (3).

1.3 Extensions

- Temporal dynamics $\partial_t x + x = h$ leads to dynamics MFT
- networks with small width; non-Gaussian corrections, can be treated dagrammatically
- fluctuations of the saddle point fields: implies pairwise correlations among the units