`osl-dynamics`: DyNeMo Cost Function

C. Gohil

OHBA, Department of Psychiatry, Warneford Hospital, Oxford, OX3 7JX

(February 11, 2023)

Abstract

We outline the derivation of the cost function used to train Dynamic Network Modes (DyNeMo) in osl-dynamics.

1 Variational Free Energy

We would like infer the underlying logit at each time point in a sequence by minimising the *variational free energy*,

$$\begin{aligned} F &= -\int ... \int q(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T) \log \left(\frac{p(\boldsymbol{x}_1, ..., \boldsymbol{x}_T | \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T) p(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T)}{q(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T)} \right) \mathrm{d}\boldsymbol{\theta}_1 ... \mathrm{d}\boldsymbol{\theta}_T, \\ &= -\int q(\boldsymbol{\theta}_{1:T}) \log \left(\frac{p(\boldsymbol{x}_{1:T} | \boldsymbol{\theta}_{1:T}) p(\boldsymbol{\theta}_{1:T})}{q(\boldsymbol{\theta}_{1:T})} \right) \mathrm{d}\boldsymbol{\theta}_{1:T}. \end{aligned} \tag{1}$$

where $q(\boldsymbol{\theta}_{1:T})$ is the *posterior*, $p(\boldsymbol{\theta}_{1:T})$ is the *prior* and $p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta}_{1:T})$ is the *likelihood*, $\boldsymbol{\theta}_i$ is the logit at each time point, $\boldsymbol{x}_i$ is the observed data at each time point, $i = 1, ..., T$ denotes the time point and we have introduced the notation $\xi_{1:T} = \xi_1, ..., \xi_T$. We can separate the logarithm into two terms,

$$\begin{aligned} F &= -\int q(\boldsymbol{\theta}_{1:T}) \log \left(p(\boldsymbol{x}_{1:T} | \boldsymbol{\theta}_{1:T}) \right) \mathrm{d}\boldsymbol{\theta}_{1:T} - \int q(\boldsymbol{\theta}_{1:T}) \log \left(\frac{q(\boldsymbol{\theta}_{1:T})}{p(\boldsymbol{\theta}_{1:T})} \right) \mathrm{d}\boldsymbol{\theta}_{1:T}. \\ &= -LL + KL. \end{aligned} \tag{2}$$

2 The Log-Likelihood Term

The first term is the *log-likelihood term*,

$$LL = \int q(\boldsymbol{\theta}_{1:T}) \log \left(p(\boldsymbol{x}_{1:T} | \boldsymbol{\theta}_{1:T}) \right) \mathrm{d}\boldsymbol{\theta}_{1:T}. \tag{3}$$

We use the mean field approximation for the posterior,

$$q(\boldsymbol{\theta}_{1:T}) = \prod_{i=1}^{T} q(\boldsymbol{\theta}_i). \tag{4}$$

Each factor $q(\boldsymbol{\theta}_i)$ is a Gaussian distribution parameterised by a mean $m_{\theta_i}(\boldsymbol{x}_{1:T})$ and variance $s^2_{\theta_i}(\boldsymbol{x}_{1:T})$, i.e.

$$q(\boldsymbol{\theta}_i) = \mathcal{N}(\boldsymbol{\theta}_i | m_{\theta_i}(\boldsymbol{x}_{1:T}), s^2_{\theta_i}(\boldsymbol{x}_{1:T})) \tag{5}$$

The parameters $m_{\theta_i}(\boldsymbol{x}_{1:T})$ and $s^2_{\theta_i}(\boldsymbol{x}_{1:T})$ are calculated using the inference RNN. We also assume the data at each time point is independent and only depends on the logit at that time point, i.e. we factorise the likelihood as

$$p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta}_{1:T}) = \prod_{j=1}^{T} p(\boldsymbol{x}_j|\boldsymbol{\theta}_j). \tag{6}$$

We assume a multivariate normal distribution for the data, i.e.

$$p(\boldsymbol{x}_j|\boldsymbol{\theta}_j) = \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{m}(\boldsymbol{\theta}_j), \boldsymbol{C}(\boldsymbol{\theta}_j)) \tag{7}$$

Substituting Eqs. (4) and (6) into Eq. (3),

$$\begin{aligned} LL &= \int \left[\prod_{i=1}^{T} q(\boldsymbol{\theta}_i)\right] \log\left(\prod_{j=1}^{T} p(\boldsymbol{x}_j|\boldsymbol{\theta}_j)\right) \mathrm{d}\boldsymbol{\theta}_{1:T} \\ &= \sum_{j=1}^{T} \int \left[\prod_{i=1}^{T} q(\boldsymbol{\theta}_i)\right] \log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j)\right) \mathrm{d}\boldsymbol{\theta}_{1:T} \end{aligned} \tag{8}$$

For each term in the summation we can factorise the integral as

$$\begin{aligned} LL &= \sum_{j=1}^{T} \left[\int q(\boldsymbol{\theta}_j) \log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j)\right) \mathrm{d}\boldsymbol{\theta}_j \prod_{i=1, i\neq j}^{T} \int q(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i\right] \\ &= \sum_{j=1}^{T} \int q(\boldsymbol{\theta}_j) \log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j)\right) \mathrm{d}\boldsymbol{\theta}_j. \end{aligned} \tag{9}$$

We can write the log-likelihood term as an expectation and use a Monte Carlo estimate to calculate it,

$$\begin{aligned} LL &= \sum_{j=1}^{T} \mathbb{E}_{q(\boldsymbol{\theta}_j)} \left\{\log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j)\right)\right\} \\ &\approx \sum_{j=1}^{T} \frac{1}{N} \sum_{s=1}^{N} \log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j^s)\right), \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}_j^s$ denotes the s^{th} sample from the posterior distribution $q(\boldsymbol{\theta}_j)$ at time point j. In practice we use just one sample, i.e. $N = 1$. Therefore, the log-likelihood term is

$$LL \approx \sum_{j=1}^{T} \log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j^1)\right). \tag{11}$$

3 The KL Term

The second term in F is the *KL divergence term*,

$$KL = \int q(\boldsymbol{\theta}_{1:T}) \log\left(\frac{q(\boldsymbol{\theta}_{1:T})}{p(\boldsymbol{\theta}_{1:T})}\right) \mathrm{d}\boldsymbol{\theta}_{1:T}. \tag{12}$$

We use the mean field approximation for the posterior (Eq. (4)) and factorise the prior as

$$p(\boldsymbol{\theta}_{1:T}) = p(\boldsymbol{\theta}_1) \prod_{k=2}^{T} p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{1:k-1}). \tag{13}$$

With these substitutions the KL divergence term becomes

$$KL = \int \left[\prod_{i=1}^{T} q(\boldsymbol{\theta}_i)\right] \log\left(\frac{\prod_{j=1}^{T} q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_1)\prod_{k=2}^{T} p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{1:k-1})}\right) \mathrm{d}\boldsymbol{\theta}_{1:T}. \tag{14}$$

We split up the logarithm as

$$KL = \int \left[\prod_{i=1}^{T} q(\boldsymbol{\theta}_i)\right] \left[\log\left(\frac{q(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_1)}\right) + \log\left(\prod_{j=2}^{T} \frac{q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})}\right)\right] \mathrm{d}\boldsymbol{\theta}_{1:T} \tag{15}$$

and clip the first logarithm to give

$$\begin{aligned} KL &\approx \int \left[\prod_{i=1}^{T} q(\boldsymbol{\theta}_i)\right] \log\left(\prod_{j=2}^{T} \frac{q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})}\right) \mathrm{d}\boldsymbol{\theta}_{1:T} \\ &\approx \sum_{j=2}^{T} \int \left[\prod_{i=1}^{T} q(\boldsymbol{\theta}_i)\right] \log\left(\frac{q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})}\right) \mathrm{d}\boldsymbol{\theta}_{1:T} \end{aligned} \tag{16}$$

For each term in the summation we can factorise the integral as

$$\begin{aligned} KL &\approx \sum_{j=2}^{T} \int \left[\prod_{i=1}^{j} q(\boldsymbol{\theta}_i)\right] \log\left(\frac{q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})}\right) \mathrm{d}\boldsymbol{\theta}_{1:j} \left[\prod_{k=j+1}^{T} \int q(\boldsymbol{\theta}_k)\mathrm{d}\boldsymbol{\theta}_k\right] \\ &\approx \sum_{j=2}^{T} \int \left[\prod_{i=1}^{j} q(\boldsymbol{\theta}_i)\right] \log\left(\frac{q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})}\right) \mathrm{d}\boldsymbol{\theta}_{1:j} \end{aligned} \tag{17}$$

We denote the integral over $\mathrm{d}\boldsymbol{\theta}_j$ by

$$D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j) \,||\, p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})\right) = \int q(\boldsymbol{\theta}_j) \log\left(\frac{q(\boldsymbol{\theta}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})}\right) \mathrm{d}\boldsymbol{\theta}_j. \tag{18}$$

Substituting this into the KL divergence term, we get

$$\begin{aligned} KL &\approx \sum_{j=2}^{T} \int \prod_{i=1}^{j-1} q(\boldsymbol{\theta}_i) D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j) \,||\, p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})\right) \mathrm{d}\boldsymbol{\theta}_{1:j-1} \\ &\approx \sum_{j=2}^{T} \mathbb{E}_{q(\boldsymbol{\theta}_1)}\left\{\mathbb{E}_{q(\boldsymbol{\theta}_2)}\left\{...\left\{\mathbb{E}_{q(\boldsymbol{\theta}_{j-1})}\left\{D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j) \,||\, p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{j-1})\right)\right\}\right\}\right\}\right\} \\ &\approx \sum_{j=2}^{T} \mathbb{E}_{q(\boldsymbol{\theta}_{1:j-1})}\left\{D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j) \mid p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1})\right)\right\} \\ &\approx \sum_{j=2}^{T} \frac{1}{N} \sum_{s=1}^{N} D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j) \,||\, p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1}^{s})\right). \end{aligned} \tag{19}$$

We use a Monte Carlo estimate of each expectation using a single sample from each posterior $q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2), ..., q(\boldsymbol{\theta}_{j-1})$. Therefore, our KL divergence term is

$$KL \approx \sum_{j=2}^{T} D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j)) \,||\, p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1}^{1})\right). \tag{20}$$

The prior $p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1}^{1})$ is a Gaussian distribution parameterised by a mean $\mu_{\theta_j}(\boldsymbol{\theta}_{1:j-1}^{1})$ and variance $\sigma_{\theta_j}^2(\boldsymbol{\theta}_{1:j-1}^{1})$, i.e.

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1}^{1}) = \mathcal{N}(\boldsymbol{\theta}_j|\mu_{\theta_j}(\boldsymbol{\theta}_{1:j-1}^{1}), \sigma_{\theta_j}^2(\boldsymbol{\theta}_{1:j-1}^{1})). \tag{21}$$

The parameters $\mu_{\theta_j}(\boldsymbol{\theta}_{1:j-1}^{1})$ and $\sigma_{\theta_j}(\boldsymbol{\theta}_{1:j-1}^{1})$ are calculated using the model RNN.

4 Cost Function

We use the variational free energy as our cost function,

$$\begin{aligned} \mathcal{L} &= F = -LL + KL \\ \mathcal{L} &= -\sum_{j=1}^{T} \log\left(p(\boldsymbol{x}_j|\boldsymbol{\theta}_j^1)\right) + \sum_{j=2}^{T} D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_j) \;||\; p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{1:j-1}^1)\right). \end{aligned} \tag{22}$$