D4 and variants

Skorik Sergey

MIPT, 2023

September 23, 2023

Backgrounds

SSM setup

Consider the problem of modeling time-series data $\boldsymbol{y}_{1:K}$, $\boldsymbol{y}_k \in \mathbb{R}^N$, where $k = 1, \ldots, K$ using dynamical latent variables $\boldsymbol{x}_{1:K}$, $\boldsymbol{x}_k \in \mathbb{R}^M$ with a Markovian property. Under the SSM modeling framework, the joint probability distribution of latent variables and observations can be factorized by conditional probabilities of a generative process defined by

$$p(\boldsymbol{x}_{1:K}, \boldsymbol{y}_{1:K}) = p(\boldsymbol{x}_1, \boldsymbol{y}_1) \cdot \prod_{t=2}^{K} p(\boldsymbol{x}_t, \boldsymbol{y}_t | \boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1})$$

$$= p(\boldsymbol{x}_1, \boldsymbol{y}_1) \cdot \prod_{t=2}^{K} p(\boldsymbol{x}_t | \boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1}) \cdot p(\boldsymbol{y}_t | \boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1})$$

Backgrounds

SSM setup

Using Markovian property

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1}) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1}) = p(\boldsymbol{y}_t|\boldsymbol{x}_t)$$

Thus, we can write

$$p(\boldsymbol{x}_{1:K}, \boldsymbol{y}_{1:K}) = p(\boldsymbol{x}_1) \cdot p(\boldsymbol{y}_1|\boldsymbol{x}_1) \cdot \prod_{t=2}^{K} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) p(\boldsymbol{y}_t|\boldsymbol{x}_t)$$

The posterior distribution is defined by the following recursive solution

$$p(\boldsymbol{x}_{1:t}, \boldsymbol{y}_{1:t}) \propto p(\boldsymbol{y}_t|\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \cdot p(\boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1})$$

D4 setup

Setup

Let's assume $\boldsymbol{h}_t = \boldsymbol{y}_{1:t-1}$, we can rewrite the posterior distribution of $\boldsymbol{x}_{1:t}$ given $\boldsymbol{y}_{1:t}$ as

$$p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_t, \boldsymbol{h}_t) = \frac{p(\boldsymbol{x}_{1:t}, \boldsymbol{y}_t, \boldsymbol{h}_t)}{p(\boldsymbol{y}_t, \boldsymbol{h}_t)} = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, \boldsymbol{h}_t)p(\boldsymbol{x}_{1:t-1}, \boldsymbol{x}_t, \boldsymbol{h}_t)}{p(\boldsymbol{y}_t, \boldsymbol{h}_t)}$$

Using Markovian assumption we can rewrite

$$p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_t, \boldsymbol{h}_t) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{h}_t)p(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})p(\boldsymbol{x}_{1:t-1}, \boldsymbol{h}_t)p(\boldsymbol{h}_t)}{p(\boldsymbol{y}_t, \boldsymbol{h}_t)}$$

Using Bayesian rule to change $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{h}_t)$ we can reduce $p(\boldsymbol{y}_t, \boldsymbol{h}_t)$ in the numerator and denominator and get

$$p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_t, \boldsymbol{h}_t) = \frac{p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{h}_t)}{p(\boldsymbol{x}_t|\boldsymbol{h}_t)}p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{1:t-1}|\boldsymbol{y}_{t-1}, \boldsymbol{h}_{t-1}) \tag{1}$$

D4 setup

D4 model

D4 is comprised of two equations

- A state transition equation

$$\boldsymbol{x}_t | \boldsymbol{x}_{t-1} \sim g(\boldsymbol{x}_{t-1}; \boldsymbol{\omega}) \tag{2}$$

- prediction process equation

$$\boldsymbol{x}_t | \boldsymbol{y}_t, \boldsymbol{h}_t \sim f(\boldsymbol{y}_t, \boldsymbol{h}_t; \Omega) \tag{3}$$

We need to integrate $p(\boldsymbol{x}_t | \boldsymbol{h}_t)$:

$$p(\boldsymbol{x}_t | \boldsymbol{h}_t) = \int p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1} | \boldsymbol{y}_{t-1}, \boldsymbol{h}_{t-1}) d\boldsymbol{x}_{t-1}$$

This can be done more efficiently using the sequential (recursive) sampling procedure: **Smoothed sequential importance sampling**

D4 model

Computational complexity

The filtering for SSM requires $\mathcal{O}(NK)$ operations to sample one trajectory of the state approximately distributed according to $p(\boldsymbol{x}_{1:K}, \boldsymbol{y}_{1:K})$. On the other hand, the D4 requires $\mathcal{O}(NK^2)$ operations to sample one path for the same distribution. But we can control $\boldsymbol{h}_t = \boldsymbol{y}_{1:t-1}$ by replacing with a fixed length history L which can reduce the computational cost to $\mathcal{O}(NKL)$.

D4 model training

EM algorithm

State variables $\boldsymbol{x}_t$ are not directly observed. For this settings we can use the EM algorithm:

$$\boldsymbol{\theta}^{r+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r) \tag{4}$$

In our case the Q is defined by:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r) = \mathbb{E}_{\boldsymbol{x}_{0:K}|\boldsymbol{y}_{1:K};\boldsymbol{\theta}^r}\left[\log\left(p(\boldsymbol{Y},\boldsymbol{X}|\boldsymbol{\theta})\right)\right] \tag{5}$$

where

$$p(\boldsymbol{Y},\boldsymbol{X}|\boldsymbol{\theta}) = p(\boldsymbol{\omega}_0;\boldsymbol{x}_0)\prod_{t=1}^{K} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1};\boldsymbol{\omega})\prod_{t=1}^{K}\frac{p(\boldsymbol{x}_t|\boldsymbol{y}_t,\boldsymbol{h}_t;\boldsymbol{\Omega})}{p(\boldsymbol{x}_t|\boldsymbol{h}_t;\boldsymbol{\Omega})}$$

and $\boldsymbol{\theta}^r = \{\boldsymbol{\omega}_0^r;\boldsymbol{\omega}^r;\boldsymbol{\Omega}^r\}$. For the simplicity of notation, we use $\mathbb{E}_K$ for $\mathbb{E}_{\boldsymbol{x}_{0:K}|\boldsymbol{y}_{1:K};\boldsymbol{\theta}^r}$.

D4 model training

E-step

We need to estimate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r)$. We can expand function, as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r) = \mathbb{E}_K[\log p(\boldsymbol{\omega}_0; \boldsymbol{x}_0) +$$

$$+\sum_{t=1}^{K} \log p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}; \boldsymbol{\omega}) + \sum_{t=1}^{K} \log p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{h}_t; \boldsymbol{\Omega})] - \mathbb{E}_K \left[\sum_{t=1}^{K} \log p(\boldsymbol{x}_t|\boldsymbol{h}_t; \boldsymbol{\Omega})\right]$$

We can bound the second term

$$\mathbb{E}_K \left[\sum_{t=1}^{K} \log p(\boldsymbol{x}_t|\boldsymbol{h}_t; \boldsymbol{\Omega})\right] = -KL\left[p(\boldsymbol{x}_t|\boldsymbol{y}_{1:K}; \boldsymbol{\theta}^r || p(\boldsymbol{x}_t|\boldsymbol{h}_t; \boldsymbol{\Omega})\right] -$$

$$-\mathbb{H}\left[p(\boldsymbol{x}_t|\boldsymbol{y}_{1:K}; \boldsymbol{\theta}^r)\right]$$

D4 model training

M-step

The following optimization problem

$$\max_{\boldsymbol{\theta}} \mathbb{E}_K [\log p(\boldsymbol{\omega}_0; \boldsymbol{x}_0) \sum_{t=1}^{K} \log p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}; \boldsymbol{\omega}) + \sum_{t=1}^{K} \log p(\boldsymbol{x}_t | \boldsymbol{y}_t, \boldsymbol{h}_t; \boldsymbol{\Omega})]$$

$$s.t. \sum_{t=1}^{K} KL\left[p(\boldsymbol{x}_t | \boldsymbol{y}_{1:K}; \boldsymbol{\theta}^r || p(\boldsymbol{x}_t | \boldsymbol{h}_t; \boldsymbol{\Omega})\right] + \mathbb{H}\left[p(\boldsymbol{x}_t | \boldsymbol{y}_{1:K}; \boldsymbol{\theta}^r)\right] < \varepsilon$$

D4 training

Algorithm 1 D4 Learning Algorithm

procedure Regularized-EM-for-D4($\boldsymbol{y}_{1:K}, \boldsymbol{\theta}^{(0)}, \lambda, D, L$)
    $\boldsymbol{h}_k \leftarrow \{\boldsymbol{y}_{k-L:k-1}\}, \boldsymbol{Q}^0 \leftarrow 0$
    Do
        $\boldsymbol{Q}^{max} \leftarrow \boldsymbol{Q}^r$
        $\tilde{\boldsymbol{x}}_{1:K}^{1:D} \leftarrow \boldsymbol{x}_{1:K}^{1:D}$
        Sample D smoothed trajectories using equations 8 and 11, $\boldsymbol{x}_{1:K}^{1:D} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:K}^{1:D} \mid \boldsymbol{y}_{1:K})$
        $\boldsymbol{\theta}^r, \boldsymbol{Q}^r = Update - Model(\boldsymbol{y}_{1:K}, \tilde{\boldsymbol{x}}_{1:K}^{1:D}, \boldsymbol{x}_{1:K}^{1:D}, \boldsymbol{\theta}^{(r-1)}, \lambda)$
    DoWhile$\{\boldsymbol{Q}^r > \boldsymbol{Q}^{max}\}$
    return $\boldsymbol{\theta}^{(r-1)}, \boldsymbol{Q}^{max}$
end procedure
procedure Update-Model($\boldsymbol{y}_{1:K}, \tilde{\boldsymbol{x}}_{1:K}^{1:D}, \boldsymbol{x}_{1:K}^{1:D}, \boldsymbol{\theta}^{(r-1)}, \lambda$)
    $\{\boldsymbol{\omega}_0^{(r-1)}, \boldsymbol{\omega}^{(r-1)}, \boldsymbol{\Omega}^{(r-1)}\} = \boldsymbol{\theta}^{(r-1)}$
    Update $\boldsymbol{Q}^r$ using equation 17 evaluate at $\boldsymbol{\theta}^{(r-1)}$
    Update $\boldsymbol{\Omega}^r$, $\boldsymbol{\omega}^r$, and $\boldsymbol{\omega}_0^r$ using gradients calculated by equations 18, 26, and 27; respectively
    return $\{\boldsymbol{\omega}_0^r, \boldsymbol{\omega}^r, \boldsymbol{\Omega}^r\}, \boldsymbol{Q}^r$
end procedure

Discussion

- The place of the article in current area
- Match notation in the article and our setup