A new information bottleneck method

April 25, 2021

1 Variational principle

For a Markov process $\{x_t\}$, we can consider the following variational problem for seeking the information bottleneck $r_t = r(x_t)$,

$$\max_{r,D_{\text{IB}}} J_{\text{IB}}(r, D_{\text{IB}}) = \mathbb{E}_t \left[\log D_{\text{IB}}\left(r_t, x_{t+\tau}\right)\right] + \mathbb{E}_{t,t'}\left[\log\left(1 - D_{\text{IB}}\left(r_t, x_{t+\tau}\right)\right)\right], \quad (1)$$

where $D_{\text{IB}}(r, x) \in (0, 1)$, $\mathbb{E}_t[\cdot]$ means expected values over all transition pairs $(x_t, x_{t+\tau})$ with lag time τ, and $\mathbb{E}_{t,t'}[\cdot]$ denotes the expectation with $x_t, x_{t'+\tau}$ are independently drawn from the trajectory.

Suppose that all tansition pairs with lag time τ are collected as $\{(x_n, y_n)\}_{n=1}^N$, we can estimate J_{IB} as

$$J_{\text{IB}}(r, D_{\text{IB}}) \approx \frac{1}{N}\sum_n \log D_{\text{IB}}\left(r\left(x_n\right), y_n\right) + \frac{1}{N}\sum_n \log\left(1 - D_{\text{IB}}\left(r\left(x_n\right), y_n\right)\right),$$

where y_n' is randomly drawn from $\{y_n\}_{n=1}^N$. Moreover, in practice, $\log D_{\text{IB}}$ and $\log(1 - D_{\text{IB}})$ can be modeled as an NN with Logsoftmax output layer.

We have the following conclusions:

Proposition 1. *For a given r,*

$$J_{\text{IB}}(r) = \max_{f_{\text{IB}}} J_{\text{IB}}(r, D_{\text{IB}}) = \text{JS}(p(r_t, x_{t+\tau})||p(r_t)p(x_{t+\tau})) + \text{const},$$

where $\text{JS}(\cdot||\cdot)$ denotes the JS divergence. Notice that the mutual information between r_t and $x_{t+\tau}$ equals $\text{KL}(p(r_t, x_{t+\tau})||p(r_t)p(x_{t+\tau}))$, so $J_{\text{IB}}(r)$ can also be interpreted as a measure of the mutual dependence between r_t and $x_{t+\tau}$.

Proposition 2. *For all mapping r,*

$$J_{\text{IB}}(r) \le J_{\text{IB}}(I_d)$$

and the equality holds if

$$p(x_{t+\tau}|r_t) = p(x_{t+\tau}|x_t),$$

where $I_d(x) = x$ denotes the identity mapping. Thus $r(x_t)$ can also be considered as a past-future information bottleneck when $J_{\text{IB}}(r)$ is maximized.

It might be more numerically stable to estimate

$$\mathbb{E}_{t,t'}\left[\log\left(1-D_{\mathrm{IB}}\left(r_t, x_{t+\tau}\right)\right)\right]$$

than to estimate

$$\log \mathbb{E}_{t,t'}\left[D_{\mathrm{IB}}\left(r_t, x_{t+\tau}\right)\right]$$

in the Donsker-Varadhan variational representation.

2 Modeling reduced kinetics

We can model the kinetics r_t as

$$\mathrm{d}r_t = -\nabla V(r_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t.$$

By selecting the step Δt, we can get the following discrete-time model for $r_t \to r_{t+\tau}$:

$$r_{t+(k+1)\Delta t} = r_{t+k\Delta t} - \nabla V(r_{t+k\Delta t})\Delta t + \sqrt{2\Delta t}u_{t+k\Delta t}, \text{ for } k = 0, \ldots, \frac{\tau}{\Delta t} - 1 \quad (2)$$

with $u_{t+k\Delta t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The model can be optimized like a GAN:

$$\min_V \max_{D_M} J_M(V, D_M) = \mathbb{E}_t\left[\log D_M\left(r_t, r_{t+\tau}\right)\right] - \mathbb{E}_t\left[\log\left(1 - D_M\left(r_t, \hat{r}_{t+\tau}\right)\right)\right],$$

where $\hat{r}_{t+\tau}$ denotes the random prediction of $r_{t+\tau}$ given by (2), and $D_M(r_t, r_{t+\tau}) \in (0,1)$. In training, $\hat{r}_{t+\tau}$ can be considered as a function of r_t and independent noise $(u_t, u_{t+\Delta t}, \ldots, u_{t+\tau-\Delta t})$ defined by (2).

3 Model reduction

By combining the previous two parts, we can perform the model reduction as

$$\min_{r,V,D_{\mathrm{IB}}} \max_{D_M} -J_{\mathrm{IB}}(r, D_{\mathrm{IB}}) + \lambda J_M(V, D_M).$$

Todo list:

1. Find the bottleneck by the new method in Section 1 without considering the model of reduced kinetics, and compare it with the current method.

2. Use the method in Section 2 to approximate the potential function of a given Brownian dynamics (e.g., Mueller potential model) without considering the dimension reduction.

3. Perform the complete model reduction by combining the results of the above two steps.