Fisher kernel and HMM gradients.

R. Huang, and C. Gohil

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

(May 2, 2023)

Abstract

We give an introduction of kernel methods and describe the use of Fisher kernel in the context of HMM. We also give derivation of the HMM variational gradients with respect to the parameters.

1 Kernel Methods

In kernel methods, instead of using the original features $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ is the space of data, for regression or classification tasks, we consider using the transformed features $\Phi(x)$, where $\Phi : \mathcal{X} \to \mathcal{H}$ is called the feature map and $\mathcal{H}$ is a usually high-dimensional if not infinte-dimensional Hilbert space. In kernel methods, such as kernel ridge regression and kernel SVM, we can avoid the evaluation of $\Phi(\boldsymbol{x})$, and only focus on the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}} \tag{1}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product on $\mathcal{H}$.

2 Fisher Kernel

We want to choose a kernel function that respect our prior knowledge of the data and is able to take into account the generative model. Fisher kernel is a reasonable choice and it is defined as

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = U_{\boldsymbol{x}_i}{}^T \mathcal{I}^{-1} U_{\boldsymbol{x}_j}, \tag{2}$$

where U_x is the Fisher score function defined by

$$U_{\boldsymbol{x}} = \nabla_w \log p(\boldsymbol{x}|w), \tag{3}$$

and $\mathcal{I}$ is the Fisher information defined by

$$\mathcal{I} = \mathbb{E}_{\boldsymbol{x}} \left[-\nabla_w^2 \log p(\boldsymbol{x}|w) \right] \tag{4}$$

Here $p(\boldsymbol{x}|w)$ is the likelihood of the data specified by the generative model and w is the parameters of the model. However, the Fisher information is either intractable or expensive to compute, the practical Fisher kernel is often used but assuming $\mathcal{I} = \boldsymbol{I}$, the identity and the resulting kernel becomes

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = U_{\boldsymbol{x}_i}{}^T U_{\boldsymbol{x}_j}. \tag{5}$$

Notice now the feature map is $\Phi : \boldsymbol{x} \mapsto U_{\boldsymbol{x}} \in l^2$.

3 Fisher Kernel for Hidden Markov Models

In `osl-dynamics`, the generative model of HMM is given by (for more details, see the write up `hmm-cost-function`)

$$p(\boldsymbol{x}_{1:T}, s_{1:T}|w) = p(\boldsymbol{x}_1|s_1, w)p(s_1) \prod_{t=2}^{T} p(\boldsymbol{x}_t|s_t, w)p(s_t|s_{t-1}), \tag{6}$$

where $w = (\theta_{\text{obs}}, \boldsymbol{A}, \boldsymbol{\pi}_1)$ are the parameters including

- the observation model parameters $\theta_{\text{obs}} = \{\boldsymbol{m}_k, \boldsymbol{C}_k\}_{k=1}^K$ where $\{\boldsymbol{m}_k\}_{k=1}^K$ and $\{\boldsymbol{C}_k\}_{k=1}^K$ are the state means and covariances respectively,
- the transition probability matrix $\boldsymbol{A}$,
- the initial state probabilities $\boldsymbol{\pi}_1$.

However, we cannot calculate the Fisher score since the likelihood of the data $p(\boldsymbol{x}_{1:T}|w)$ under this generative model is intractable and has no analytic solution. Hence we instead use the gradients of the variational free energy with respect to the parameters w. The variational free energy of our HMM has analytic form as

$$\begin{aligned}
\mathcal{F}(\boldsymbol{x}_{1:T}, w) &= -LL + E + P \\
&= -\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \log\left[\mathcal{N}(\boldsymbol{x}_t|\boldsymbol{m}_k, \boldsymbol{C}_k)\right] \\
&+ \sum_{t=1}^{T-1}\sum_{i,j=1}^{K} \xi_{ij}(t)\log\xi_{ij}(t) - \sum_{t=2}^{T-1}\sum_{i=1}^{K}\gamma_i(t)\log\gamma_i(t) \\
&+ \sum_{i=1}^{K}\gamma_i(1)\log\pi_{1,i} + \sum_{t=1}^{T-1}\sum_{i,j=1}^{K}\xi_{ij}(t)\log A_{ij}.
\end{aligned} \tag{7}$$

4 Derivative of the Variational Free Energy

To use the Fisher Kernel, we need to compute the derivative of the variational free energy with respect to the parameters $\theta_{\text{obs}}, \boldsymbol{\pi}_1, \boldsymbol{A}$.

4.1 Derivative with respect to $\boldsymbol{\pi}_1$

The posterior expected prior probability P is the only term in the free energy that depends on $\boldsymbol{\pi}_1$,

$$\begin{aligned}
\nabla_{\boldsymbol{\pi}_1}\mathcal{F} &= \nabla_{\boldsymbol{\pi}_1}\sum_{i=1}^{K}\gamma_i(1)\log\pi_{1,i} = \nabla_{\boldsymbol{\pi}_1}\gamma(1)^T\log\boldsymbol{\pi}_1 \\
&= \frac{\gamma(1)}{\boldsymbol{\pi}_1}.
\end{aligned} \tag{8}$$

4.2 Derivative with respect to $\boldsymbol{A}$

The posterior expected prior probability P is also the only term that depends on $\boldsymbol{A}$,

$$\begin{aligned}
\nabla_{\boldsymbol{A}}\mathcal{F} &= \nabla_{\boldsymbol{A}}\sum_{t=1}^{T-1}\sum_{i,j=1}^{K}\xi_{ij}(t)\log A_{ij} \\
&= \sum_{t=1}^{T-1}\frac{\boldsymbol{\xi}(t)}{\boldsymbol{A}} = \frac{\sum_{t=1}^{T-1}\boldsymbol{\xi}(t)}{\boldsymbol{A}}.
\end{aligned} \tag{9}$$

4.3 Derivative with respect to the Observation Model Parameters

The only term depending on the observation model parameters $\{m_l, C_l\}_{l=1}^{K}$ is the posterior expected log likelihood. In `osl-dynamics`, we are already using gradient-based methods to train the observation model parameters and hence it is straight forward to compute the gradients with respect to the means and covariances (using tensorflow gradient tape).