Supplementary Material

Christian Wildner and Heinz Koeppl

1. Introduction

This document contains supplementary material for the main article ”Moment Based Variational Inference for Markov Jump Processes”. The section structure of this document mimics the structure of the main document.

2. Background

2.1. Decomposition of the Divergence

Here, we show that the divergence between an approximate process and a posterior process can be decomposed into the divergence of the approximate process and the prior process plus a contribution of of the observations. More specifically, we prove Eq. (6) of the main text.

Lemma 1. *The divergence in Eq. (6) of the main tex can be written as*

$$D\left[P^Z \,||\, P^{\bar{X}}\right] = D\left[P^Z \,||\, P^X\right] - \sum_{k=1}^{n} \mathsf{E}[\log p(y_k \mid Z(t_k))] + \text{const}$$

and hence the variational problem is independent of the exact functional form of the posterior intensities.

To see this, let $Z(t)$ be a MJP with marginal distribution p^Z and time-dependent transition function λ. To simplify the derivation, it is useful to introduce operators $\mathcal{L}, \mathcal{L}^\dagger$ defined by their action on a function f of a suitable class

$$\begin{aligned} \mathcal{L}f(x) &= \sum_{y \neq x} \lambda(y, x, t) f(y) - \sum_{y \neq x} \lambda(x, y, t) f(x)\,, \\ \mathcal{L}^\dagger f(x) &= \sum_{y \neq x} (f(y) - f(x)) \lambda(x, y, t)\,. \end{aligned} \tag{1}$$

Using Eq. (1) of the main text for a function $g : \mathcal{X} \times [0, T] \to \mathbb{R}$ can be written as

$$\frac{d}{dt}\mathsf{E}[g(Z(t), t)] = \mathsf{E}[\mathcal{L}^\dagger g(Z(t), t)] + \mathsf{E}[\partial_t g(Z(t), t)] \tag{2}$$

where the second expression comes from the explicit time dependence of the function g. Note that (2) can be obtained by inserting the distribution function p^Z, using the product rule of differentiation and inserting the master equation obeyed by p^Z. Following the notation of the main text, we will denote the time independent transition function of the prior process by Q. Starting from the divergence of two time dependent MJPs (Eq. (2), main text) and inserting the posterior intensities yields

$$D\left[P^Z \,||\, P^{\bar{X}}\right] = J_1 - J_2 + J_3 + J_4 \tag{3}$$

with

$$
\begin{aligned}
J_1 &= \sum_y \int_0^T \mathsf{E}\left[\frac{\sigma(y,t)}{\sigma(Z(t),t)} Q(Z(t),y)\right] dt\,,\\
J_2 &= \sum_y \int_0^T \mathsf{E}\left[\lambda(Z(t),y,t)\right] dt\\
J_3 &= \sum_y \int_0^T \mathsf{E}\left[\lambda(Z(t),y,t)\log\left(\lambda(Z(t),y,t)\right)\right] dt\\
J_4 &= \sum_y \int_0^T \mathsf{E}\left[\lambda(Z(t),y,t)\log\left(\frac{\sigma(Z(t),t)}{\sigma(y,t)}\right)\right] dt
\end{aligned}
$$

Now observe that by (1) the expectation within J_4 can be written as

$$\sum_y \mathsf{E}\left[\lambda(Z(t),y,t)\log\left(\frac{\sigma(Z(t),t)}{\sigma(y,t)}\right)\right] = -\mathsf{E}[\mathcal{L}^\dagger \log\sigma(Z(t),t)]\,. \tag{4}$$

By addition of a zero, we can rewrite the term J_1 as

$$J_1 = \sum_y \int_0^T \mathsf{E}\left[\frac{\sigma(y,t)-\sigma(Z_i(t),t)}{\sigma(Z(t),t)} Q(Z(t),y)\right] dt + \sum_y \int_0^T \mathsf{E}\left[Q(Z(t),y)\right] dt\,. \tag{5}$$

The term in the numerator within the first expectation corresponds to the left hand side of the backward equation obeyed by σ (Eq. 4, main text). Inserting this and using $\frac{d}{dx}\log f(x) = \frac{1}{f(x)}\frac{d}{dx}f(x)$ leads to

$$J_1 = -\int_0^T \mathsf{E}[\partial_t \log\sigma(Z(t),t)]dt + \sum_y \int_0^T \mathsf{E}\left[Q(Z(t),y)\right] dt\,. \tag{6}$$

Inserting the relations (6) and (4) into (3) and exploiting (2) we get

$$D\left[P^Z \,||\, P^{\bar{X}}\right] = D\left[P^Z \,||\, P^X\right] - \int_0^T \frac{d}{dt}\mathsf{E}[\log\sigma(Z(t),t)]dt\,.$$

Due to the jump conditions of the backward equation, the integral on the right hand side evaluates to

$$
\begin{aligned}
\int_0^T \frac{d}{dt}\mathsf{E}[\log\sigma(Z(t),t)]dt = {}& \mathsf{E}[\log\sigma(Z(T),T)] - \mathsf{E}[\log\sigma(Z(0),0)]\\
&+ \sum_{k=1}^n \left(\mathsf{E}[\log\sigma(Z(t_k^-),t_k^-)] - \mathsf{E}[\log\sigma(Z(t_k),t_k)]\right)
\end{aligned}
$$

The terms in the first line correspond to the contributions of the terminal and the initial time and can be ignored. For the terminal time, this follows directly from the terminal constraint $\sigma(x,T)=1$ for all x. The contribution of the initial term, we observe that by definition of σ we have

$$\mathsf{E}[\log\sigma(Z(0),0)] = \mathsf{E}[\log p(y_1,\ldots,y_n \mid Z(0))] = \log p(y_1,\ldots,y_n \mid Z(0)) =: \log Z\,,$$

which is constant with respect to the variational transition function λ since the initial distribution is fixed. Note that in the usual language of variational inference $\log Z$ corresponds to the marginal log likelihood of the data. Finally, exploiting the reset conditions of the backward equation, we get

$$\int_0^T \frac{d}{dt}\mathsf{E}[\log\sigma(Z(t),t)]dt = \sum_{k=1}^n \mathsf{E}[\log p(y_k \mid Z(t_k))]\,.$$

In summary, we get

$$D\left[P^Z \,||\, P^{\bar{X}}\right] = D\left[P^Z \,||\, P^X\right] - \sum_{k=1}^n \mathsf{E}[\log p(y_k \mid Z(t_k))] + \log Z$$

which is in line with the usual decomposition of the KL divergence into the evidence $\log Z$ and a free energy contribution $L = \sum_{k=1}^n \mathsf{E}[\log p(y_k \mid Z(t_k))] - D\left[P^Z \,||\, P^X\right]$ (Blei et al., 2017).

3. Moment Based Variational Smoothing

3.1. Maximum Principle

Our goal is to solve the variational problem in the form of the control problem as given in Eq. (13) of the main text. However, here we consider a slightly more general scenario in which we do not solve for the natural moments φ directly. Instead, we choose a collection of moment function ψ such that the natural moments can be represented by

$$\varphi(t) = g(\psi(t)) \tag{7}$$

for a suitable map g. Note that we can always find such a collection of moments by choosing $\psi = \varphi$ and $g = \mathrm{Id}$. In this formulation, control problem (20) of the main text becomes

$$\begin{aligned} &\text{minimize} && L[\lambda, \psi] - F[\psi] \\ &\text{subject to} && \frac{d}{dt}\psi(t) = f(\lambda(t), \psi(t))] \end{aligned} \tag{8}$$

where we assume the initial conditions as fixed and known. We follow the indirect approach known from optimal control and variational calculus by introducing the Langrange multiplier functions (or co-states) η_i, $i = 1, \dots, r$ to enforce the ODE relation between ψ and λ. We obtain the Lagrangian functional

$$J[\lambda, \psi, \eta] = L[\lambda, \psi] - F[\psi] - \int_0^T \eta(t)^T \left[f(\lambda(t), \psi(t) - \dot{\psi}(t) \right] . \tag{9}$$

where we have stacked the η_i into a single vector η. Since the functional F only acts on the discrete observation times, we will ignore it for a moment. Computing the functional derivative with respect to ψ and setting it to zero leads to the co-state equations

$$\frac{d}{dt}\eta_i(t) = \sum_{j=1}^{r} \frac{dg_j}{d\psi_i} \left(1 - \lambda_j(t) + \lambda_j(t) \log \lambda_j(t)\right) - \sum_{j=1}^{r} \frac{df_j}{d\psi_i} \eta_j(t) \tag{10}$$

valid in between the observations. At the point of the observations, the functional F will induce jump conditions for η given by

$$\lim_{t \nearrow t_k} \eta_i(t) = \eta_i(t_k) + \frac{d}{d\psi_i(t_k)} \mathsf{E}[p(y_k \mid Z(t_k))] . \tag{11}$$

It is therefore crucial that we express the expected log likelihood with respect to the variational process in terms of the moment functions ψ. If this is not naturally possible, we may enlarge the space of moment functions suitably. Next, consider the functional derivatives with respect to λ_i leading to

$$0 = g_i(\psi(t)) \log \lambda_i(t) - \sum_{j=1}^{r} \eta_j(t) \frac{df_j}{d\lambda_i} . \tag{12}$$

The equations (10), (11), (12) together with the ODE for ψ from a set of necessary conditions for optimal solutions known as Pontryagin's maximum (minimum) principle. Note that since the function f is typically linear in λ, it may be possible to solve (12) for λ and eliminate it in (10). In general, the benefit of this questionable because the resulting ODE for η is highly non-linear.

3.2. Gradient Based Optimization

A simple approach to solve to obtain a numerical solution from the maximum principle is the forward backward sweep (Mcasey et al., 2012). Here, one starts with an initial guess for λ and solves the forward equation. The forward solution is then used to solve (10) backward in time. We may then solve (12) for λ to obtain an update given the forward and backward solution. Iterating this procedure may lead to a stationary point of the functional.

In our applications, this procedure turned out to be highly unstable, probably because the updates are too large in the geometry of the probabilistic manifold defined by λ. We therefore modified the procedure as follows. We keep the forward an backward solution steps, but instead of solving (12) for λ, we understand the r.h.s of (12) as the gradient of the functional L when considered as a function of λ alone. More explicitly, we use

$$(\nabla L[\lambda])_i = g_i(\psi(t)) \log \lambda_i(t) - \sum_{j=1}^{r} \eta_j(t) \frac{df_j}{d\lambda_i} . \tag{13}$$

Algorithm 1 Basic Gradient Descent for MBVI

```
Input: Initial guess for the scaling factors $\lambda^{(0)}(t)$,
initial condition $\psi(0)$.
repeat
    Given $\lambda^{(n)}(t)$ and $\psi(0)$, compute $\psi^{(n)}(t)$ using (8).
    Given $\lambda^{(n)}(t)$ and $\psi^{(n)}(t)$, compute $\eta^{(n)}(t)$ using (10), (11).
    Compute current gradient $\nabla L[\lambda^{(n)}]$ according to (13).
    Compute $\lambda^{(n+1)}$ from (14).
until $|L[\lambda^{(n)}] - L[\lambda^{(n-1)}]| < \text{tolerance}$
Output: Optimized variational scaling factor $\lambda^*$ .
```

Using (13) allows to apply a gradient descent type algorithm in the scaling factors λ by using update steps of the form

$$\lambda^{(n+1)} = \lambda^{(n)} - h\nabla L[\lambda^{(n)}] \tag{14}$$

where h is the step size. An algorithmic description of this procedure in form of pseudo code is given in Alg. 1.

3.3. Natural Gradient

It is well-known that gradient-based algorithms can perform poorly on manifolds. One solution to this is to incorporate the local geometry of the manifold via its metric tensor G. Now consider a family of probability distributions p_θ parametrized by θ. Then the set of all θ spans a manifold whose geometric structure is given by the Fisher information matrix (Amari, 1998). This defines the so called natural gradient

$$\tilde{\nabla} p_\theta = G^{-1} \nabla p_\theta \,.$$

Discrete update steps using the natural gradient corresponds to an approximate steepest descent with respect to the local geometry of p_θ. In order to transfer this setting, we exploit a connection between the KL divergence and the fisher information metric

$$D[p_\theta \,||\, p_{\theta'}] = \frac{1}{2}(\theta - \theta')^T \, G(\theta)(\theta - \theta') + o((\theta' - \theta)^2) \,,$$

that is the metric G arises from second order expansion of the KL divergence in the parameter of the second argument. While the Fisher information matrix requires a finite dimensional parameter θ, the second order expansion of the KL divergence can be transferred to the path space setting. Thus, consider the variational family corresponding to a partition Π of the transition space and consider variational scaling factors λ, λ'. Then the divergence of Z^λ and $Z^{\lambda'}$ as in Section 3.1 of the main text can be written as

$$D\left[P^\lambda \,||\, P^{\lambda'}\right] = \int_0^T \varphi_i(t) \left(\lambda_i'(t) - \lambda_i(t) + \lambda_i(t) \frac{\lambda_i(t)}{\lambda_i'(t)}\right) \mathrm{d}t$$

where the φ_i depend on λ as they are defined as expected values with respect to Z^λ. Now rewrite the logarithmic term as

$$\log\left(\frac{\lambda_i'(t)}{\lambda_i(t)}\right) = \log\left(1 + \frac{\lambda_i'(t) - \lambda_i(t)}{\lambda_i(t)}\right) \,. \tag{15}$$

If $\sup_{t\in[0,T]} |\lambda'(t) - \lambda(t)|$ is small, we may use the standard approximation

$$\log(1+x) = x - \frac{x^2}{2} + O(x^3) \,. \tag{16}$$

Applying (16) to (15) and inserting the result into (3.3) causes the linear terms to cancel and we are left with

$$D\left[P^\lambda \,||\, P^{\lambda'}\right] = \frac{1}{2}\int_0^T \sum_{i=1}^R \frac{\varphi_i(t))}{\lambda_i(t)} (\lambda_i'(t) - \lambda_i(t))^2 \mathrm{d}t + \int_0^T O((\lambda_i'(t) - \lambda_i(t))^3) \mathrm{d}t \,. \tag{17}$$

We can understand the above expression as the infinitesimal distance between the path distributions corresponding to the variational parameters λ and λ' for a fixed partition. From the second order expansion (17), in combination with (12), we get the natural gradient

$$\tilde{\nabla} L[\lambda] = \lambda_i(t) \log \lambda_i(t) - \frac{\lambda_i(t)}{\varphi_i(t)} \sum_{j=1}^{r} \eta_j(t) \frac{df_j}{d\lambda_i} \,. \tag{18}$$

where $\varphi = g(\psi)$ and η are evaluated for the current value of λ via the forward and backward equations. To perform the optimization, we simply have to replace the gradient evaluation in Alg. 1 by (18). Doing so not only increased speed and reliability of the optimization but also led to a visually smoother transition from the prior to the posterior in all considered examples.

4. Parameter Inference

4.1. Statistics for Expectation Maximization

By the modified definition of the φ_i as

$$\lambda(x, y, t) = \lambda_i(t) h(x, y) \quad \text{for} \quad (x, y) \in \Pi_i \,,$$

The functional L changes slightly to

$$L[\varphi, \lambda, \theta] = \sum_{i=1}^{r} \int_0^T \varphi_i(t) \left(c_i(\theta) - \lambda_i(t) + \lambda_i(t) \log \frac{\lambda_i(t)}{c_i(\theta)} \right) dt \,. \tag{19}$$

By breaking (19) down into individual terms, it becomes clear how the summary statistics

$$\begin{aligned} G_i &= \int_0^T \varphi_i(t) dt \,, \\ H_i &= \int_0^T \varphi_i(t) \lambda_i(t) dt \end{aligned} \tag{20}$$

arise. Considering L as a function of θ for fixed λ and φ, we may write

$$L[\theta] = \sum_{i=1}^{r} \left(G_i c_i(\theta) - H_i \log c_i(\theta) \right) + const \,. \tag{21}$$

From the last expression, we obtain a stationarity condition by differentiating with respect to θ.

4.2. Bayesian Approach

Consider a general scenario where with a hierarchical model given by

$$p(\theta, x, y) = p(\theta) p(x \mid \theta) p(y \mid x) \tag{22}$$

where we aim to approximate the joint posterior $p(\theta, x \mid y)$ by a product $q(\theta) q(x)$. It is straightforward to show that the optimal parameter posterior $q^*(\theta)$ satisfies

$$q^*(\theta) \propto p(c) \exp \left(-D[q(x) \, || p(x \mid \theta) p(y \mid x)] \right) \,. \tag{23}$$

Transferred to our setting, the expression within the exponential becomes the functional L and since we only care about the parts depending on θ, we may as well insert (21) leading to Eq. (22) of the main text.

5. Examples

Here we provide explicit forms of the variational functionals of the different examples and the corresponding stationarity conditions.

5.1. Gene Expression Model

For the gene expression model, the functions φ_i can be expressed in terms of the first order expectations

$$m_i(t) = \mathsf{E}[Z_i(t)]\,.$$

Due to the Gaussian observation model, we also require second order moments

$$m_{ij}(t) = \mathsf{E}[(Z_i(t)Z_j(t))]\,.$$

It is therefore convenient to parametrize the dynamical system in these terms. For the first order moments, we obtain the equations

$$\begin{aligned}
\frac{d}{dt}m_1(t) &= \lambda_1(t)(1 - m_1(t)) - \lambda_2(t)m_1(t)\,,\\
\frac{d}{dt}m_2(t) &= \lambda_3(t)m_1(t) - \lambda_4(t)m_2(t)\,,\\
\frac{d}{dt}m_3(t))\rangle &= \lambda_5 m_2(t) - \lambda_6(t)m_3(t)\,.
\end{aligned}$$

For three species, we get 6 additional second order equations

$$\begin{aligned}
\frac{d}{dt}m_{11}(t) &= \lambda_1(t)m_1(t) + \lambda_1(t) - 2\lambda_1(t)m_{11}(t) - 2\lambda_2 m_{11}(t) + \lambda_2(t)m_1(t)\,,\\
\frac{d}{dt}m_{12}(t) &= \lambda_1(t)m_2(t) - \lambda_1(t)m_{12}(t) - \lambda_2(t)m_{12}(t) + \lambda_3(t)m_{11} - \lambda_4 m_{12}(t)\,,\\
\frac{d}{dt}m_{13}(t) &= \lambda_1(t)m_3(t) - \lambda_1(t)m_{13}(t) - \lambda_2(t)m_{13}(t) + \lambda_5(t)m_{12}(t) - \lambda_6(t)m_{13}(t)\,,\\
\frac{d}{dt}m_{22}(t) &= 2\lambda_3(t)m_{12}(t) + \lambda_3(t)m_1(t) - 2\lambda_4(t)m_{22}(t) + \lambda_4(t)m_2(t)\,,\\
\frac{d}{dt}m_{23}(t) &= \lambda_3(t)m_{13}(t) - \lambda_4(t)m_{23}(t) + \lambda_5(t)m_{22}(t)\lambda_6(t)m_{23}(t)\,,\\
\frac{d}{dt}m_{33}(t) &= 2\lambda_5(t)m_{23}(t) + \lambda_5(t)m_2(t) - 2\lambda_6(t)m_{33}(t) + \lambda_6(t)m_3(t)\,.
\end{aligned}$$

5.2. Predator Prey Model

As before, we take into account the first and second order moments and choose a corresponding parametrization. Suppressing the explicit time arguments for the sake of readability, the resulting system is given by

$$\begin{aligned}
\dot{m}_1 &= \lambda_1 m_1 - \lambda_2\left(m_{12} + m_1 m_2\right)\,,\\
\dot{m}_2 &= \lambda_3\left(m_{12} + m_1 m_2\right) - \lambda_4 m_2\,,\\
\dot{m}_{11} &= 2\lambda_1 m_{11} + \lambda_1 m_1 + \lambda_2(m_{12} + m_1 m_2) + 2\lambda_2 m_1\left(m_{12} + m_1 m_2\right) - 2\lambda_2 m_{112}\,,\\
\dot{m}_{12} &= \lambda_1 m_{12} + \lambda_2(m_{12} + m_1 m_2)m_2 - \lambda_3(m_{12} + m_1 m_2)m_1 - \lambda_4 m_{12} - \lambda_2 m_{122} + \lambda_3 m_{112}\,,\\
\dot{m}_{22} &= -2\lambda_3 m_2(m_{12} + m_1 m_2) + \lambda_3(m_{12} + m_1 m_2) - 2\lambda_4 m_{22} + \lambda_4 m_2 + 2\lambda_3 m_{122}\,.
\end{aligned}$$

As is typical for systems with non-linear intensity functions, the moment equation up to order three are not closed but depend on the (non-central) third-order moments m_{112} and m_{122}. In order to obtain a finite dimensional system, we have to use a moment closure method that expresses the higher order moments in terms of lower order moments

$$\begin{aligned}
m_{112} &= V_1(m_1, m_2, m_{11}, m_{12}, m_{22})\,,\\
m_{122} &= V_2(m_1, m_2, m_{11}, m_{12}, m_{22})\,.
\end{aligned}$$

While many standard moment closure approaches use ad hoc choices for the closure functions V_1, V_2, we follow the recently proposed variational moment closure approach (Bronstein & Koeppl, 2018) that allows for a systematic derivation of closure functions based on a set of moment functions and a distributional ansatz. In particular we use a log-normal product Poisson mixture distribution and get

$$\mathsf{E}[X_1^2 X_2] = \frac{\left(\mathsf{E}[X_1^2] - \mathsf{E}[X_1]\right)\mathsf{E}[X_1 X_2]^2}{\mathsf{E}[X_1]^2 \mathsf{E}[X_2]} + \mathsf{E}[X_1 X_2] \tag{24}$$

and a similar expression for $\mathsf{E}[X_1 X_2^2]$.

References

Amari, S.-I. Natural Gradient Works Efficiently in Learning. *Neural Comput.*, 10(2):251–276, 1998.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bronstein, L. and Koeppl, H. A variational approach to moment-closure approximations for the kinetics of biomolecular reaction networks. *The Journal of Chemical Physics*, 148(1):014105, 2018.

Mcasey, M., Mou, L., and Han, W. Convergence of the forward-backward sweep method in optimal control. *Computational Optimization and Applications*, 53, 09 2012.