Variational Inference for the PLDS

Lars Buesing

March 6, 2017

1 Model

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mu, \Sigma) && (1) \\
p(\mathbf{y}|\mathbf{x}) &= \prod_n p(y_n|\eta_n), \qquad \eta := W\mathbf{x} + \overline{\mathbf{d}} && (2)
\end{aligned}
$$

For latent dynamical system we know that $\Lambda := \Sigma^{-1}$ is tri-diagonal, and W block-diagonal:

$$
\begin{aligned}
\Lambda &= \begin{pmatrix} Q_0^{-1} + AQ^{-1}A^\top & A^\top Q^{-1} & & \\ Q^{-1}A & Q^{-1} + AQ^{-1}A^\top & A^\top Q^{-1} & \\ & \ddots & \ddots & \ddots \end{pmatrix} && (3) \\
W &= \text{blk-diag}(\underbrace{C, \ldots, C}_{T\text{-times}}) && (4)
\end{aligned}
$$

2 Inference problem

Gaussian variational approximation:

$$
q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, V) \tag{5}
$$

Variational lower bound:

$$
\begin{aligned}
\mathcal{L}(\mathbf{m}, V) &\leq \log p(\mathbf{y}) && (6) \\
\mathcal{L}(\mathbf{m}, V) &= \frac{1}{2}\big(\log|V| - \text{tr}[\Sigma^{-1}V] - (\mathbf{m} - \mu)^\top \Sigma^{-1}(\mathbf{m} - \mu)\big) + \sum_n \mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)] && (7)
\end{aligned}
$$

For Poisson with exp link function we can compute $\mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)]$, otherwise (eg for Bernoulli observations) use a local variational lower bound on the integrated likelihood:

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{x})}[\log p(y_n|\eta_n)] &\geq -f_n(\overline{m}_n, \overline{v}_n) && (8) \\
f_n(\overline{m}_n, \overline{v}_n) &= -y_n\overline{m}_n + \exp(\overline{m}_n + \overline{v}_n/2) && (9) \\
\overline{\mathbf{m}} &:= W\mathbf{m} + \overline{\mathbf{d}} && (10) \\
\overline{\mathbf{v}} &:= \text{diag}(WVW^\top) && (11)
\end{aligned}
$$

The bound then reads:

$$
\mathcal{L}(\mathbf{m}, V) = \frac{1}{2}\big(\log|V| - \text{tr}[\Sigma^{-1}V] - (\mathbf{m} - \mu)^\top \Sigma^{-1}(\mathbf{m} - \mu)\big) - \sum_n f_n(\overline{m}_n, \overline{v}_n) \tag{12}
$$

For convex f_n (true for exp-PLDS): strictly concave optimization in $\mathbf{m}, V$
Possible optimization strategies:

1. Direct optimization over $\mathbf{m}, V$: strictly concave, however V dense; does not make use of Markovian structure of the model
2. Optimization over $\mathbf{m}, V^{-1}$: Opper et al show that optimal $V^* = (\Sigma^{-1} + W^\top \text{diag}(\lambda)W)^{-1}$; hence for tri-diagonal Σ^{-1} and block-diagonal W then V^* is also tri-diagonal; however optimization over $\mathbf{m}, \lambda$ is not convex and converges slowly according to [Seeger et al. ICML2013]
3. Solve the dual optimization as proposed in [Seeger et al. ICML2013]: convex, makes use of Markovian structure of the model

3 Variational inference via dual optimization

3.1 Optimization to solve

Dual problem:

$$\begin{aligned}\min_{\lambda} \quad & \frac{1}{2}(\lambda-\mathbf{y})^\top W\Sigma W^\top(\lambda-\mathbf{y}) - (W\mu+\overline{\mathbf{d}})^\top(\lambda-\mathbf{y}) - \frac{1}{2}\log|A_\lambda| + \sum_n f^*(\lambda_n) \\ \text{subject to} \quad & \lambda_i > 0\end{aligned} \tag{13}$$

where

$$f^*(\lambda_i) := \lambda_i(\log\lambda_i - 1) \tag{14}$$

$$A_\lambda := \Sigma^{-1} + W^\top \operatorname{diag}(\lambda) W \tag{15}$$

The optimal variational parameters for $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}^*, V^*)$ are given by:

$$\mathbf{m}^* = \mu - \Sigma W^\top(\lambda^* - \mathbf{y}) \tag{16}$$

$$V^* = (\Sigma^{-1} + W^\top \operatorname{diag}(\lambda^*)W)^{-1} = A_{\lambda^*}^{-1} \tag{17}$$

3.2 How to optimize?

Use gradient based methods:

$$\begin{aligned}\nabla_\lambda &= W\Sigma W^\top(\lambda-\mathbf{y}) - W\mu - \overline{\mathbf{d}} + \log\lambda - \frac{1}{2}\operatorname{diag}(WA_\lambda^{-1}W^\top) \\ &= \underbrace{W\Sigma W\lambda}_{O(N)} + \underbrace{\log\lambda}_{O(N)} - \frac{1}{2}\operatorname{diag}(\underbrace{WA_\lambda^{-1}W^\top}_{O(N)}) - \underbrace{W(\Sigma W^\top\mathbf{y}+\mu)}_{\text{pre-compute}}\end{aligned}$$

Hessian:

$$H_\lambda = \operatorname{diag}(\lambda)^{-1} + W\Sigma W^\top + (WA_\lambda^{-1}W^\top)\circ(WA_\lambda^{-1}W^\top)$$

Iterate:

$$\begin{aligned}\mathbf{m}^k &= \mu + \Sigma W^\top\mathbf{y} - \Sigma W^\top\lambda^k \\ A^k &= \Sigma^{-1} + W^\top\operatorname{diag}(\lambda^k)W \\ \nabla^k &= \log\lambda^k - W\mathbf{m}^k - \overline{\mathbf{d}} - \frac{1}{2}\operatorname{diag}(W(A^k)^{-1}W^\top) \\ \lambda^{k+1} &= \lambda^k - \nu\nabla^k\end{aligned}$$

Computing the block-diagonal elements of A^k is equivalent to Kalman smoothing and requires a forward-backward pass through the data which costs $O(Td^3)$.
What's the relation to Laplace approximation?

$$\begin{aligned}\nabla^k &= -\Sigma^{-1}(\mathbf{x}-\mu) + W^\top(\mathbf{y} - \exp(W\mathbf{x}+\overline{\mathbf{d}})) \\ H^k &= -(\Sigma^{-1} + W^\top\operatorname{diag}(\exp(W\mathbf{x}+\overline{\mathbf{d}}))W)\end{aligned}$$

3.3 Kalman smoothing

The matrix A_λ equals exactly the precision matrix of a LDS with dynamics given by A, Q and observations sampled from $\mathcal{N}(C\mathbf{x}_t, \operatorname{diag}(\lambda_t))$. Hence, calculating the block-diagonal of A_λ^{-1} is exactly equivalent to calculating the smoothed posterior covariance of this LDS. Let $P_{t|t}$ denote the filtered covariance, $P_{t+1|t}$ the one-step-ahead covariance and $P_{t|T}$ the smoothed covariance of this model.

$$(A_\lambda^{-1})_{(t-1)d+1:td,(t-1)d+1:td} \overset{!}{=} P_{t|T} \tag{18}$$

We use the Kalman smoother recursions. The forward pass reads:

$$
\begin{aligned}
P_{t+1|t} &= A P_{t|t} A^\top + Q && (19)\\
P_{t+1|t+1} &= \left(P_{t+1|t}^{-1} + C^\top \operatorname{diag}(\lambda_t) C\right)^{-1} && (20)\\
&= \left(I + P_{t+1|t} C^\top \operatorname{diag}(\lambda_t) C\right) \backslash P_{t+1|t} && (21)\\
P_{0|0} &= Q_0 && (22)
\end{aligned}
$$

The backward pass is given by:

$$
\begin{aligned}
C_t &= P_{t|t} A^\top / P_{t+1|t} && (23)\\
P_{t|T} &= P_{t|t} + C_t \left(P_{t+1|T} - P_{t+1|t}\right) C_t^\top && (24)
\end{aligned}
$$

The initialization for the backward pass $P_{T|T}$ is calculated the last step of the forward pass.

4 Appendix

4.1 Derivation of dual optimization

Original primal problem:

$$\begin{array}{ll} \max\limits_{\mathbf{m},V} & \frac{1}{2}\left(\log|V| - \mathrm{tr}[\Sigma^{-1}V] - \|\mathbf{m}-\mu\|^2_{\Sigma^{-1}}\right) - \sum_n f_n(\overline{m}_n, \overline{v}_n) \\ \text{s.t.} & V \in S^{++} \end{array}$$

Expanded primal problem:

$$\begin{array}{ll} \mathrm{argmax}_{\mathbf{m},V,\rho,h} & \frac{1}{2}\left(\log|V| - \mathrm{tr}[\Sigma^{-1}V] - \|\mathbf{m}-\mu\|^2_{\Sigma^{-1}}\right) - \sum_n f_n(h_n, \rho_n) \\ \text{s.t.} & V \in S^{++} \\ & h = W\mathbf{m} + \overline{\mathbf{d}} \\ & \rho = \mathrm{diag}(WVW^\top) \end{array}$$

Lagrangian:

$$\begin{aligned} \mathcal{L}(\mathbf{m}, V, h, \rho, \alpha, \lambda) \quad := \quad & \frac{1}{2}\left(\log|V| - \mathrm{tr}[\Sigma^{-1}V] - \|\mathbf{m}-\mu\|^2_{\Sigma^{-1}}\right) - \sum_n f_n(h_n, \rho_n) \\ & + \alpha^\top(h - W\mathbf{m} + \overline{\mathbf{d}}) + \frac{1}{2}\lambda^\top(\rho - \mathrm{diag}(WVW^\top)) \end{aligned}$$

Dual

$$\begin{aligned} D(\alpha, \lambda) \quad &:= \quad \min_{\mathbf{m},V,h,\rho} L(\mathbf{m}, V, h, \rho, \alpha, \lambda) \\ V^* \quad &= \quad (\Sigma^{-1} + W^\top \mathrm{diag}(\lambda) W)^{-1} =: A_\lambda^{-1} \\ \mathbf{m}^* \quad &= \quad \mu - \Sigma W^\top \alpha \\ \alpha \quad &= \quad \lambda - y \end{aligned}$$

Final reduced dual problem:

$$\begin{array}{ll} \mathrm{argmin}_\lambda & \frac{1}{2}(\lambda - \mathbf{y})^\top W \Sigma W (\lambda - \mathbf{y}) - (\overline{\mathbf{d}} + W\mu)^\top(\lambda - \mathbf{y}) - \frac{1}{2}\log|A_\lambda| + \sum_n f^*(\lambda_n) \\ \text{s.t.} & \lambda_i > 0 \end{array}$$

4.2 Duality basics

Primal problem with optimal values p^*:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & f_i(x) \leq 0 \\ & h_i(x) = 0 \end{array}$$

Lagrange function with $\lambda_i \geq 0$:

$$L(x, \lambda, \nu) \quad := \quad f(x) + \sum_i \lambda_i f_i(x) + \sum_i \nu_i h_i(x)$$

Dual:

$$g(\lambda, \nu) \quad := \quad \inf_x L(x, \lambda, \nu)$$

Dual is a lower bound:

$$g(\lambda, \nu) \quad \leq \quad p^*$$

This can be shown by bounding the constraint functions with linear functions from below. Dual problem:

$$\begin{array}{ll} \min & -g(\lambda, \nu) \\ \text{s.t.} & \lambda_i \geq 0 \end{array}$$

Dual is always convex!

Slater's conditions We have stong duality iff:

$$g(\lambda, \nu) \quad = \quad p^*$$

A sufficient condition for strong duality is: f convex, no inequality constraints, primal feasible

Dual function The dual f^* of a function f is defined as:

$$f^*(y) \quad := \quad \sup_x \; y^\top x - f(x)$$