Variational Stabilized Linear Forgetting in State-Space Models (Supplementary Material)

Thijs van de Laar[*], Marco Cox[*], Anouk van Diepen[*], and Bert de Vries[*,†]
[*]Department of Electrical Engineering, Eindhoven University of Technology
[†]GN Hearing, Eindhoven, the Netherlands

February 7, 2017

This document details the derivations of the non-standard variational messages in [1]. Variational message updates are computed in accordance with [2]. We introduce shorthand notations for expectations over recognition distributions $q(\cdot)$ with respect to a function $f(\cdot)$: for the mean $\overline{f(\cdot)} = \mathbb{E}_{(\cdot)}[f(\cdot)] = \mathbb{E}_{q(\cdot)}[f(\cdot)]$; covariance $\mathbb{C}\mathrm{ov}[f(\cdot)] = \mathbb{C}\mathrm{ov}_{q(\cdot)}[f(\cdot)]$, and average energy $\mathbb{U}[f(\cdot)] = -\mathbb{E}_{q(\cdot)}[\log f(\cdot)]$. Variational messages are denoted by a function ν. Throughout the derivations, terms independent of the message argument are absorbed in a constant C.

Bernoulli node

The Bernoulli node models a Bernoulli distribution over a switch $z \in \{0, 1\}$, as parametrized by $\pi \in [0, 1]$. The Bernoulli node, with node function $f(z, \pi) = \pi^z (1 - \pi)^{1-z}$, is drawn in Fig. 1:

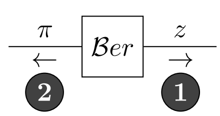

Figure 1: Factor graph representation of a Bernoulli node.

Update for message 1

$$
\begin{aligned}
\log \nu(z) &= \mathbb{E}_\pi[\log f(z,\pi)] + C \\
&= \mathbb{E}_\pi\left[\log\left(\pi^z (1-\pi)^{1-z}\right)\right] + C \\
&= z\mathbb{E}_\pi[\log \pi] + (1-z)\mathbb{E}_\pi[\log(1-\pi)] + C \\
\nu(z) &\propto \exp\left(\overline{\log \pi}\right)^z \times \exp\left(\overline{\log(1-\pi)}\right)^{1-z} \; .
\end{aligned}
$$

After renormalizing the event probabilities, we obtain a Bernoulli distribution:

$$
\nu(z) \propto \mathcal{B}er\left(z | \frac{\exp\left(\overline{\log \pi}\right)}{\exp\left(\overline{\log \pi}\right) + \exp\left(\overline{\log(1-\pi)}\right)}\right) \; .
$$

As a specific example, when we assume $q(\pi) = \mathcal{B}eta(\pi|a,b)$ beta distributed, the variational update becomes:

$$
\begin{aligned}
\nu(z) &\propto \exp(\psi(a) - \psi(a+b))^z \times \exp(\psi(b) - \psi(a+b))^{1-z} \\
&= \exp(-\psi(a+b)) \times \exp(\psi(a))^z \times \exp(\psi(b))^{1-z} \\
&\propto \mathcal{B}er\left(z | \frac{\exp(\psi(a))}{\exp(\psi(a)) + \exp(\psi(b))}\right) \; ,
\end{aligned}
$$

with ψ the digamma function.

Update for message 2

$$
\begin{aligned}
\log \nu(\pi) &= \mathbb{E}_z[\log f(z,\pi)] + C \\
&= \mathbb{E}_z\left[\log\left(\pi^z (1-\pi)^{1-z}\right)\right] + C \\
&= \bar{z} \log \pi + (1-\bar{z}) \log(1-\pi) + C \\
\nu(\pi) &\propto \pi^{\bar{z}} (1-\pi)^{1-\bar{z}} \\
&\propto \mathcal{B}eta(\pi | \bar{z}+1, 2-\bar{z}) \; .
\end{aligned}
$$

Gaussian mixture node

The Gaussian mixture node governs the mixing of Gaussian components, as regulated by a discrete switch $z \in \{0,1\}$. For stabilized linear forgetting, the Gaussian mixture has two Gaussian components. Fig. 2 depicts the composite Gaussian mixture node for two components, with node function:

$$
f(\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\right)^z \times \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_2, \mathbf{W}_2^{-1}\right)^{1-z} \; .
$$

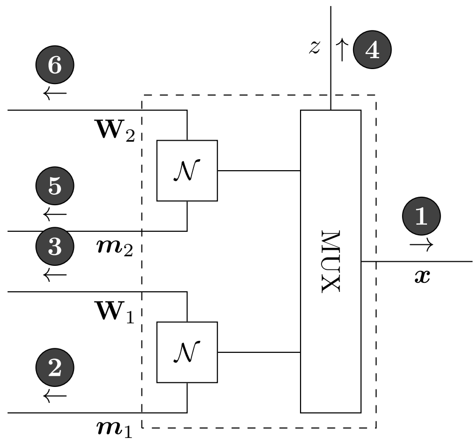

Figure 2: Factor graph representation of a Gaussian mixture composite node with two components.

Update for message 1

$$
\begin{aligned}
\log \nu(\boldsymbol{x}) &= \underset{\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2}{\mathbb{E}} [\log f(\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x})] + C \\
&= \underset{\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2}{\mathbb{E}} \Big[\log \Big(\mathcal{N}\big(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\big)^z \times \mathcal{N}\big(\boldsymbol{x} | \boldsymbol{m}_2, \mathbf{W}_2^{-1}\big)^{1-z} \Big) \Big] + C \\
&= -\frac{\bar{z}}{2} \underset{\boldsymbol{m}_1, \mathbf{W}_1}{\mathbb{E}} \big[(\boldsymbol{x} - \boldsymbol{m}_1)^T \mathbf{W}_1 (\boldsymbol{x} - \boldsymbol{m}_1) \big] - \frac{1 - \bar{z}}{2} \underset{\boldsymbol{m}_2, \mathbf{W}_2}{\mathbb{E}} \big[(\boldsymbol{x} - \boldsymbol{m})^T \mathbf{W}_2 (\boldsymbol{x} - \boldsymbol{m}_2) \big] + C \\
&= -\frac{\bar{z}}{2} (\boldsymbol{x} - \bar{\boldsymbol{m}}_1)^T \bar{\mathbf{W}}_1 (\boldsymbol{x} - \bar{\boldsymbol{m}}_1) - \frac{1 - \bar{z}}{2} (\boldsymbol{x} - \bar{\boldsymbol{m}}_2)^T \bar{\mathbf{W}}_2 (\boldsymbol{x} - \bar{\boldsymbol{m}}_2) + C \\
\nu(\boldsymbol{x}) &\propto \mathcal{N}\big(\boldsymbol{x} | \bar{\boldsymbol{m}}_1, (\bar{z} \bar{\mathbf{W}}_1)^{-1}\big) \times \mathcal{N}\big(\boldsymbol{x} | \bar{\boldsymbol{m}}_2, ((1 - \bar{z}) \bar{\mathbf{W}}_2)^{-1}\big) \ ,
\end{aligned}
$$

which yields another Gaussian distribution.

Update for message 2

$$
\begin{aligned}
\log \nu(\boldsymbol{m}_1) &= \underset{\mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x}}{\mathbb{E}} [\log f(\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x})] + C \\
&= \underset{\mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x}}{\mathbb{E}} \Big[\log \Big(\mathcal{N}\big(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\big)^z \times \mathcal{N}\big(\boldsymbol{x} | \boldsymbol{m}_2, \mathbf{W}_2^{-1}\big)^{1-z} \Big) \Big] + C \\
&= \bar{z} \underset{\mathbf{W}_1, \boldsymbol{x}}{\mathbb{E}} \big[\log \mathcal{N}\big(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\big) \big] + C \\
\nu(\boldsymbol{m}_1) &\propto \mathcal{N}\big(\boldsymbol{m}_1 | \bar{\boldsymbol{x}}, (\bar{z} \bar{\mathbf{W}}_1)^{-1}\big) \ .
\end{aligned}
$$

Update for message ③

$$
\begin{aligned}
\log \nu(\mathbf{W}_1) &= \mathop{\mathbb{E}}_{\boldsymbol{m}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x}}[\log f(\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x})] + C \\
&= \bar{z} \mathop{\mathbb{E}}_{\boldsymbol{m}_1, \boldsymbol{x}}\left[\log \mathcal{N}\left(\boldsymbol{m}_1 | \boldsymbol{x}, \mathbf{W}_1^{-1}\right)\right] + C \\
&= \frac{\bar{z}}{2}\left(\log |\mathbf{W}_1| - \mathop{\mathbb{E}}_{\boldsymbol{m}_1, \boldsymbol{x}}\left[(\boldsymbol{m}_1 - \boldsymbol{x})^T \mathbf{W}_1 (\boldsymbol{m}_1 - \boldsymbol{x})\right]\right) + C \\
\nu(\mathbf{W}_1) &\propto \log\left(|\mathbf{W}_1|^{\frac{\bar{z}}{2}}\right) - \frac{1}{2}\operatorname{tr}\left(\mathbf{W}_1 \bar{z}\left[(\bar{\boldsymbol{m}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{m}}_1 - \bar{\boldsymbol{x}})^T + \mathbb{C}\mathrm{ov}[\boldsymbol{m}_1] + \mathbb{C}\mathrm{ov}[\boldsymbol{x}]\right]\right) \,,
\end{aligned}
$$

which we recognize as a Wishart distribution:

$$
\nu(\mathbf{W}_1) \propto \mathcal{W}(\mathbf{W}_1 | \mathbf{V}, n) \,,
$$

with:

$$
\begin{aligned}
n &= \bar{z} + d + 1\,; \\
\mathbf{V} &= \left(\bar{z}\left[(\bar{\boldsymbol{m}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{m}}_1 - \bar{\boldsymbol{x}})^T + \mathbb{C}\mathrm{ov}[\boldsymbol{m}_1] + \mathbb{C}\mathrm{ov}[\boldsymbol{x}]\right]\right)^{-1} \,,
\end{aligned}
$$

with dimensionality d.

Update for message ④

$$
\begin{aligned}
\log \nu(z) &= \mathop{\mathbb{E}}_{\boldsymbol{m}_1, \mathbf{W}_1, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x}}[\log f(\boldsymbol{m}_1, \mathbf{W}_1, z, \boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x})] + C \\
&= z \mathop{\mathbb{E}}_{\boldsymbol{m}_1, \mathbf{W}_1, \boldsymbol{x}}\left[\log \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\right)\right] + (1 - z) \mathop{\mathbb{E}}_{\boldsymbol{m}_2, \mathbf{W}_2, \boldsymbol{x}}\left[\log \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_2, \mathbf{W}_2^{-1}\right)\right] + C \\
&= -z \mathbb{U}\left[\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\right)\right] - (1 - z) \mathbb{U}\left[\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_2, \mathbf{W}_2^{-1}\right)\right] + C\,,
\end{aligned}
$$

with $\mathbb{U}$ the average energy functional,

$$
\nu(z) \propto \underbrace{\exp\left(-\mathbb{U}\left[\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_1, \mathbf{W}_1^{-1}\right)\right]\right)}_{\rho_1}{}^{z} \times \underbrace{\exp\left(-\mathbb{U}\left[\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{m}_2, \mathbf{W}_2^{-1}\right)\right]\right)}_{\rho_2}{}^{1-z} \,.
$$

After renormalizing the event probabilities, we obtain a Bernoulli distribution:

$$
\nu(z) \propto \mathcal{B}er\left(z | \frac{\rho_1}{\rho_1 + \rho_2}\right) \,.
$$

Update for message ⑤

The update for $\nu(\boldsymbol{m}_2)$ is derived analogous to $\nu(\boldsymbol{m}_1)$.

$$
\nu(\boldsymbol{m}_2) \propto \mathcal{N}\left(\boldsymbol{m}_2 | \bar{\boldsymbol{x}}, \left((1 - \bar{z})\bar{\mathbf{W}}_2\right)^{-1}\right) \,.
$$

Update for message 6

The update for $\nu(\mathbf{W}_2)$ is derived analogous to $\nu(\mathbf{W}_1)$.

$$\nu(\mathbf{W}_2) \propto \mathcal{W}(\mathbf{W}_2|\mathbf{V}, n) \ ,$$

with:

$$\begin{aligned} n &= 2 - \bar{z} + d\,; \\ \mathbf{V} &= \Big((1-\bar{z})\,\big[\,(\bar{\boldsymbol{m}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{m}}_2 - \bar{\boldsymbol{x}})^T + \mathbb{C}\mathrm{ov}[\boldsymbol{m}_2] + \mathbb{C}\mathrm{ov}[\boldsymbol{x}]\,\big]\Big)^{-1} \ . \end{aligned}$$

References

[1] T. van de Laar, M. Cox, A. van Diepen, and B. de Vries, "Variational stabilized linear forgetting in state-space models," 2017, manuscript submitted for publication.

[2] J. Dauwels, "On Variational Message Passing on Factor Graphs," in *IEEE International Symposium on Information Theory*, Jun. 2007, pp. 2546–2550.