On diagonal approximations to the extended Kalman filter for online training of Bayesian neural networks

Author Name1 ABC@SAMPLE.COM
Address 1

Author Name2 XYZ@SAMPLE.COM
Address 2

Abstract

We present two approaches to approximate online Bayesian inference for the parameters of DNNs. Both are based on diagonal Gaussian approximations and linearize the network at each step to ensure efficient computation. The first approach optimizes the exclusive KL, $D_{\mathbb{KL}}(q \parallel p)$; this amounts to matching the marginal mean and *precision* of p and q. The second approach optimizes the inclusive KL, $D_{\mathbb{KL}}(q \parallel p)$, which amounts to matching the marginal mean and *variance* of p and q. The latter approach turns out to be equivalent to the previously proposed "fully decoupled EKF" approach. We show experimentally that inclusive KL yields better results than exclusive KL, and both are more effective than one-pass SGD.

Keywords: Bayesian inference, variational inference, online learning, extended Kalman filter, deep neural networks, non-stationary distributions, continual learning.

1. Introduction

In this short paper, we consider the problem of online Bayesian inference for the parameters of a neural network. That is, we want to recursively compute $p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}, \mathbf{x}_{1:t})$ in an efficient manner, where $\mathbf{x}_t \in \mathbb{R}^{N_x}$ are the input features, $\mathbf{y}_t \in \mathbb{R}^{N_y}$ are the output labels, and $\boldsymbol{\theta}_t \in \mathbb{R}^{N_z}$ are the model parameters. We assume the likelihood has the form $p(\mathbf{y}_t|\boldsymbol{\theta}_t, \mathbf{x}_t) = \text{expfam}(\mathbf{y}_t|g^{-1}(h(\mathbf{x}_t, \boldsymbol{\theta}_t)))$, where h is a neural network that predicts the natural parameters $\boldsymbol{\eta}$ of the exponential family output, and g is some link function that maps from natural parameters $\boldsymbol{\eta}$ to moment parameters $\mathbf{m}$. To handle non-stationary distributions, we allow the parameters to drift according to a Gaussian random walk, so $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = N(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{Q}_t)$, where $\mathbf{Q}_t = \text{diag}(\mathbf{q}_t)$ is the process noise. For continual learning problems where a single task drifts over time, this method tracks the optimal parameters and their uncertainty. If instead we assume $\mathbf{q}_t = 0$, so the parameters are stationary, then the method can also be used to compute $p(\boldsymbol{\theta}|\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$ with a single pass over the data.

Exact Bayesian inference in this model family is intractable, so we must resort to approximations. In the signal processing and engineering communities, a standard approach (e.g., Singhal and Wu, 1989; Puskorius and Feldkamp, 2003) is the extended Kalman filter (EKF), which amounts to linearizing the observation model h at each step, and then performing an exact Bayesian update. In the machine learning literature, it is more common (e.g., Challis and Barber, 2013; Broderick et al., 2013; Nguyen et al., 2018; Lambert et al.,

© 2022 A. Name1 & A. Name2.

2021b) to solve the following variational inference problem, using exclusive KL divergence:

$$q_t = \arg\min_q D_{\mathbb{KL}}(q(\boldsymbol{\theta}_t) \parallel p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t})) \approx \arg\min_q D_{\mathbb{KL}}(q(\boldsymbol{\theta}_t) \parallel p(\mathbf{y}_t|\boldsymbol{\theta}_t) q_{t|t-1}(\boldsymbol{\theta}_t)) \tag{1}$$

where we omit the conditioning on $\mathbf{x}_t$ for brevity, and where the one-step-ahead predictive distribution is given by

$$q_{t|t-1}(\boldsymbol{\theta}_t) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) q_{t-1}(\boldsymbol{\theta}_{t-1}) d\boldsymbol{\theta}_t = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \tag{2}$$

Some prior work (e.g., Hernández-Lobato and Adams, 2015; Soudry et al., 2014; Ghosh et al., 2016) instead uses the assumed density filtering (ADF) approach, which uses the inclusive KL divergence:

$$q_t = \arg\min_q D_{\mathbb{KL}}(p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) \parallel q(\boldsymbol{\theta}_t)) \tag{3}$$

In this paper, we focus our attention on (block) diagonal Gaussian approximations of the form $q_t(\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where $\boldsymbol{\Sigma}_t = \text{diag}(\boldsymbol{\Sigma}_t^{ii})$. Unfortunately, even with this assumption, optimizing these KL objectives is challenging, and often requires Monte Carlo approximations. We show how we can use the EKF approximation to compute the optimum of both (approximate) KL objectives in closed form in linear time. When we apply this approach to the exclusive KL objective, the resulting algorithm is (as far as we know) a novel variational approach to DNNs. When we apply this approach to the inclusive KL objective, we recover the "decoupled EKF" method (Puskorius and Feldkamp, 1991, 2003; Murtuza and Chorian, 1994).[1]

In Section 2, we describe our inference methods in more detail, and in Section 3, we perform an experimental comparison. We show that the inclusive KL objective gives better results than the exclusive KL, and both are more statistically efficient than stochastic gradient descent (SGD), which requires multiple passes over the data, making it harder to apply to the streaming setting. (We leave a comparison to online SGD variants that use replay buffers, such as Hu et al. (2021), to future work.) We conclude in Section 4.

2. Methods

In this section, we briefly describe our methods. For the derivation, please see the appendix. (We use the notation $\mathbf{A}_i$ and $\mathbf{A}^i$ to represent the ith row and column of a matrix $\mathbf{A}$, and $\mathbf{A}^{ii}$ to denote the ith (block) diagonal entry, if $\mathbf{A}$ is square.)

2.1. Extended Kalman filter

For completeness, we summarize the usual EKF procedure, specialized to our setting. For notational simplicity, we drop the conditioning on inputs $\mathbf{x}_t$.

The predict step is given by Eq. (2). If we make a diagonal assumption on the process noise, we have $\boldsymbol{\mu}_{t|t-1} = \boldsymbol{\mu}_{t-1}$ and $\boldsymbol{\Sigma}_{t|t-1} = \text{diag}(\boldsymbol{\Sigma}_{t-1}^{ii}) + \text{diag}(\mathbf{Q}_t^{ii})$.

1. If we generalize the EKF + inclusive KL update to the offline setting, using repeated forward-backward passes, as in expectation propagation, we can derive a diagonal version of the method in Kamthe et al. (2022); however, in this paper, we focus on the online setting.

If we assume a Gaussian likelihood, $p(\mathbf{y}_t|\boldsymbol{\theta}_t) = \mathcal{N}(\mathbf{y}_t|h(\boldsymbol{\theta}_t), \mathbf{R}_t)$, then the update step becomes $p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) = N(\boldsymbol{\theta}_t|\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$ where

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - h(\boldsymbol{\mu}_{t|t-1})) \tag{4}$$

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1} = \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\mathsf{T}\mathbf{R}_t^{-1}\mathbf{H}_t\right)^{-1} \tag{5}$$

$$\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\mathsf{T}\left(\mathbf{R}_t + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\mathsf{T}\right)^{-1} = \boldsymbol{\Sigma}_{t|t}\mathbf{H}_t^\mathsf{T}\mathbf{R}_t^{-1} \tag{6}$$

where $\mathbf{H}_t \equiv \mathrm{Jac}(\mathbf{h})(\boldsymbol{\mu}_{t|t-1})$ is a linearization of the observation model. (The EKF approach can be extended to exponential family likelihoods using the techniques from Ollivier (2018); Tronarp et al. (2018).) In block form, we can write the elements of the posterior covariance matrix as follows:

$$\boldsymbol{\Sigma}_{t|t}^{ij} = \boldsymbol{\Sigma}_{t|t-1}^{ij} - \boldsymbol{\Sigma}_{t|t-1}^{ii}(\mathbf{H}_t^i)^\mathsf{T}\mathbf{S}_t^{-1}\mathbf{H}_t^j\boldsymbol{\Sigma}_{t|t-1}^{jj} \tag{7}$$

$$\mathbf{S}_t = \mathbf{R}_t + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\mathsf{T} = \mathbf{R}_t + \sum_j \mathbf{H}_t^j\boldsymbol{\Sigma}_{t|t-1}^{jj}(\mathbf{H}_t^j)^\mathsf{T} \tag{8}$$

2.2. Approximating the update step

Let $p = p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ be the full covariance computed by EKF, where $\boldsymbol{\mu}_p = \boldsymbol{\mu}_{t|t}$ and $\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_{t|t}$, as in Section 2.1. In general the EKF takes $O(N_z^2)$ time to compute. We seek to approximate this by $q = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, where $\boldsymbol{\Sigma}_q = \mathrm{diag}(\boldsymbol{\Sigma}^{ii})$. We discuss different objectives below, which can be optimized in $O(N_z)$ time.

2.3. Exclusive KL (variational diagonal EKF)

Suppose we want to minimize $D_{\mathbb{KL}}(q \parallel p)$, which is mode seeking, using diagonal q. We call this the "variational diagonal EKF" or "VD-EKF" method. In the appendix we show that the result is given by $\boldsymbol{\mu}_q = \boldsymbol{\mu}_p$ and

$$(\boldsymbol{\Sigma}_{t|t}^{ii})^{-1} = (\boldsymbol{\Sigma}_{t|t}^{-1})^{ii} = \left(\boldsymbol{\Sigma}_{t|t-1}^{ii}\right)^{-1} + (\mathbf{H}_t^i)^\mathsf{T}\mathbf{R}_t^{-1}\mathbf{H}_t^i \tag{9}$$

This should be compared to the exact update in Eq. (15). We see that the variational approximation requires computing the diagonal elements of the posterior *precision matrix*, $\boldsymbol{\Sigma}_{t|t}^{-1}$. By using the matrix inversion lemma, we can also write the result as follows:

$$\boldsymbol{\Sigma}_{t|t}^{ii} = \boldsymbol{\Sigma}_{t|t-1}^{ii} - \boldsymbol{\Sigma}_{t|t-1}^{ii}(\mathbf{H}_t^i)^\mathsf{T}\left(\mathbf{R}_t + \mathbf{H}_t^i\boldsymbol{\Sigma}_{t|t-1}^{ii}(\mathbf{H}_t^i)^\mathsf{T}\right)^{-1}\mathbf{H}_t^i\boldsymbol{\Sigma}_{t|t-1}^{ii} \tag{10}$$

This is similar to the diagonal elements in the update of Eq. (7), but drops some of the terms in $\mathbf{S}_t$; we discuss the implications of this below.

2.4. Inclusive KL (fully decoupled EKF)

We now seek to minimize $D_{\mathbb{KL}}(p \parallel q)$, which is mode covering, using diagonal q. In the appendix we show that the result is given by $\boldsymbol{\mu}_q = \boldsymbol{\mu}_p$ and

$$\boldsymbol{\Sigma}_{t|t}^{ii} = \boldsymbol{\Sigma}_{t|t-1}^{ii} - \boldsymbol{\Sigma}_{t|t-1}^{ii}(\mathbf{H}_t^i)^\mathsf{T}\left(\mathbf{R}_t + \sum_j \mathbf{H}_t^j\boldsymbol{\Sigma}_{t|t-1}^{jj}(\mathbf{H}_t^j)^\mathsf{T}\right)^{-1}\mathbf{H}_t^i\boldsymbol{\Sigma}_{t|t-1}^{ii} \tag{11}$$

This requires computing the diagonal elements of the posterior *covariance matrix*, $\mathbf{\Sigma}_{t|t}$. The resulting update is equivalent to the fully decoupled EKF method of Puskorius and Feldkamp (2003); we therefore call this the "FD-EKF" method.

2.5. Discussion

We see that the main difference between the exclusive KL in Eq. (10) and the inclusive KL in Eq. (11) is that the former replaces $\mathbf{S}_t$ with $\mathbf{S}_t^i := \mathbf{R}_t + \mathbf{H}_t^i \mathbf{\Sigma}_{t|t-1}^{ii} (\mathbf{H}_t^i)^\mathsf{T}$. Intuitively, using $\mathbf{S}_t$ should be superior to $\mathbf{S}_t^i$, since $\mathbf{S}_t$ captures more of the uncertainty. In particular, from the perspective of estimating $\boldsymbol{\theta}^i$, the missing terms in $\mathbf{S}_t^i$, namely $\sum_{j \neq i} \mathbf{H}_t^j \mathbf{\Sigma}_{t|t-1}^{jj} (\mathbf{H}_t^j)^\mathsf{T}$, play the same role as the observation covariance $\mathbf{R}_t$, but reflect uncertainty in the other parameters. In Section 3, we will show experimentally that the inclusive KL gives better results.

3. Experimental results

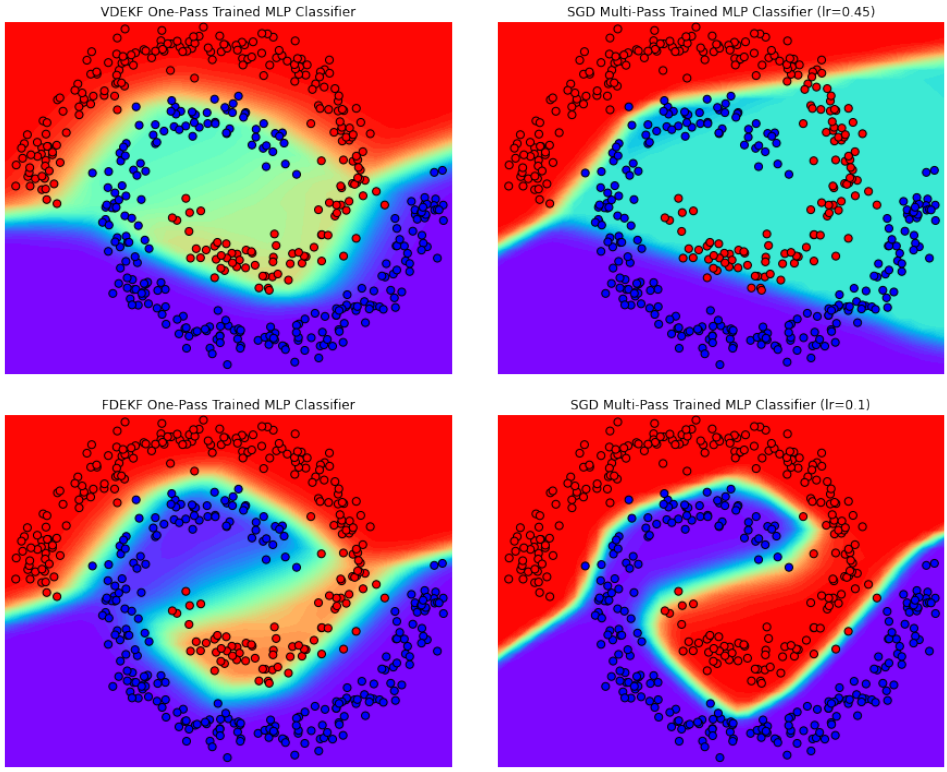

Figure 1: Comparison of posterior predictions of VD-EKF (top left), SGD with high learning rate (top right), FD-EKF (bottom left), and SGD with low learning rate (bottom right). Results are from a two-layer MLP fit to 500 data samples.

Since the exclusive KL objective neglects some forms of parameter uncertainty, it "trusts" the likelihood more, which is similar to using SGD with a larger learning rate. This is illus-

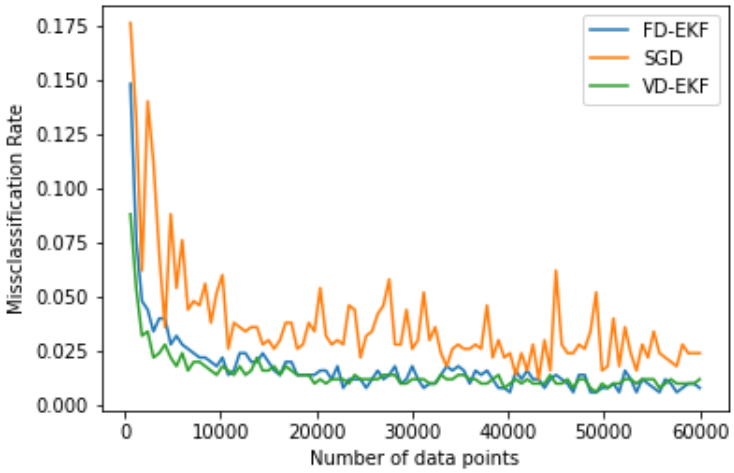

Figure 2: Error rates vs number of daa points for SGD vs VDEKF and FDEKF on spiral dataset.

trated in Fig. 1, where we display the posterior predictive distribution from a small MLP fit to a 2d binary classification dataset. We see that one pass of VD-EKF gives similar results to 5 epochs of SGD with a large learning rate of $\eta = 0.45$, whereas one pass of FD-EKF gives similar results to 5 epochs of SGD with a much smaller learning rate of $\eta = 0.1$.

In Fig. 2, we show that both VD-EKF and FD-EKF can be much more data efficient than SGD, in the sense that they need to see less data to converge to a good predictive distribution.

4. Conclusion and future work

We have seen how optimizing the inclusive KL, as is common in ADF and expectation propagation, can give better results than optimizing the exclusive KL, which is the basis of variational inference. We have also shown how we can efficiently compute both forms of update by linearizing the likelihood model, as in the EKF.

In the future, we would like to extend this approach to include more expressive approximations, such as diagonal plus low rank (e.g., see Mishkin et al., 2018; Lambert et al., 2021a). We also want to apply it to more complex non-stationary and continual learning problems, which may be achievable with more sophisticated dynamical priors.

References

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. In *NIPS*, 2013. URL http://arxiv.org/abs/1307.6769.

E Challis and D Barber. Gaussian Kullback-Leibler approximate inference. *JMLR*, 14:2239–2286, 2013. URL https://www.jmlr.org/papers/volume14/challis13a/challis13a.pdf.

Soumya Ghosh, Francesco Maria Delle Fave, and Jonathan Yedidia. Assumed density filtering methods for learning bayesian neural networks. In *AAAI*, 2016. URL https://jonathanyedidia.files.wordpress.com/2012/01/assumeddensityfilteringaaai2016final.pdf.

José Miguel Hernández-Lobato and Ryan P Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *ICML*, 2015. URL http://arxiv.org/abs/1502.05336.

Huiyi Hu, Ang Li, Daniele Calandriello, and Dilan Gorur. One pass ImageNet. In *NeurIPS 2021 Workshop on Imagenet: past, present and future*, November 2021. URL http://arxiv.org/abs/2111.01956.

Sanket Kamthe, So Takao, Shakir Mohamed, and Marc Peter Deisenroth. Iterative state estimation in non-linear dynamical systems using approximate expectation propagation. *Trans. on Machine Learning Research*, 2022. URL https://openreview.net/pdf?id=xyt4wfdo4J.

Marc Lambert, Silvère Bonnabel, and Francis Bach. The limited-memory recursive variational gaussian approximation (L-RVGA). December 2021a. URL https://hal.inria.fr/hal-03501920.

Marc Lambert, Silvère Bonnabel, and Francis Bach. The recursive variational gaussian approximation (R-VGA). *Stat. Comput.*, 32(1):10, December 2021b. URL https://hal.inria.fr/hal-03086627/document.

Aaron Mishkin, Frederik Kunstner, Didrik Nielsen, Mark Schmidt, and Mohammad Emtiyaz Khan. SLANG: Fast structured covariance approximations for bayesian deep learning with natural gradient. In *NIPS*, pages 6245–6255. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7862-slang-fast-structured-covariance-approximations-for-bayesian-deep-learning-with-natu pdf.

S Murtuza and S F Chorian. Node decoupled extended kalman filter based learning algorithm for neural networks. In *Proceedings of 1994 9th IEEE International Symposium on Intelligent Control*, pages 364–369, August 1994. URL http://dx.doi.org/10.1109/ISIC.1994.367790.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *ICLR*, 2018. URL https://openreview.net/forum?id=BkQqq0gRb.

Yann Ollivier. Online natural gradient as a kalman filter. *Electron. J. Stat.*, 12(2):2930–2961, 2018. URL https://projecteuclid.org/euclid.ejs/1537257630.

G V Puskorius and L A Feldkamp. Decoupled extended kalman filter training of feedforward layered networks. In *International Joint Conference on Neural Networks*, volume i, pages 771–777 vol.1, 1991. URL http://dx.doi.org/10.1109/IJCNN.1991.155276.

Gintaras V Puskorius and Lee A Feldkamp. Parameter-based kalman filter training: Theory and implementation. In Simon Haykin, editor, *Kalman Filtering and Neural Networks*, pages 23–67. John Wiley & Sons, Inc., 2003. URL https://onlinelibrary.wiley.com/doi/10.1002/0471221546.ch2.

Sharad Singhal and Lance Wu. Training multilayer perceptrons with the extended kalman algorithm. In *NIPS*, volume 1, 1989. URL https://proceedings.neurips.cc/paper/1988/file/38b3eff8baf56627478ec76a704e9b52-Paper.pdf.

Daniel Soudry, Itay Hubara, and Ron Meir. Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights. In *NIPS*, 2014. URL https://papers.nips.cc/paper/2014/file/076a0c97d09cf1a0ec3e19c7f2529f2b-Paper.pdf.

Filip Tronarp, Ángel F García-Fernández, and Simo Särkkä. Iterative filtering and smoothing in nonlinear and Non-Gaussian systems using conditional moments. *IEEE Signal Process. Lett.*, 25(3):408–412, 2018. URL https://acris.aalto.fi/ws/portalfiles/portal/17669270/cm_parapub.pdf.

Appendix A. Appendix

In this section, we give the derivation of the algorithms.

A.1. Variational diagonal EKF

A.1.1. Covariance

The covariance update for the EKF corresponds to the following equation:

$$\mathbf{\Sigma}_{t|t} = \mathbf{\Sigma}_{t|t-1} - \mathbf{K}_t \mathbf{H}_t \mathbf{\Sigma}_{t|t-1} \tag{12}$$

$$= \mathbf{\Sigma}_{t|t-1} - \mathbf{\Sigma}_{t|t-1} \mathbf{H}_t^\intercal \left(\mathbf{R}_t + \mathbf{H}_t \mathbf{\Sigma}_{t|t-1} \mathbf{H}_t^\intercal\right)^{-1} \mathbf{H}_t \mathbf{\Sigma}_{t|t-1} \tag{13}$$

The **Woodbury matrix inversion lemma** states the following:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{14}$$

Letting $A = \mathbf{\Sigma}_{t|t-1}^{-1}, U = \mathbf{H}_t^\intercal, C = \mathbf{R}_t^{-1}, V = \mathbf{H}_t$ and applying the inversion lemma leads to the following simplified covariance update:

$$\mathbf{\Sigma}_{t|t} = \left(\mathbf{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\intercal \mathbf{R}_t^{-1} \mathbf{H}_t\right)^{-1} \tag{15}$$

We make the variational assumption that the covariance is fully diagonal at each time step. The prior is assumed to be:

$$\mathbf{\Sigma}_{t|t-1} \approx \operatorname{diag}\left(\left(\sigma_{t|t-1}^i\right)^2\right) \tag{16}$$

and we would like a diagonal posterior $\boldsymbol{\Sigma}_{t|t}$ that is the closest approximation to the expression on the right of Eq. (15), i.e., one that minimizes the KL-divergence between the Gaussians.

The KL-divergence between D-dimensional Gaussians $p \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $q \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ can be simplified to the following expression:

$$D_{\mathbb{KL}}(q \parallel p) = \frac{1}{2}\left(\log \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - D + \left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p\right)^{\intercal} \boldsymbol{\Sigma}_p \left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p\right) + \operatorname{tr}\left\{\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\right\}\right) \tag{17}$$

For the means, the optimal solution is to set $\hat{\boldsymbol{\mu}}_q = \boldsymbol{\mu}_p$. We can verify this by setting the partial derivative of $D_{\mathbb{KL}}(q \parallel p)$ with respect to the mean elements $\mu_{q,i}$ to zero.

Since the covariance and the mean contribute additively to the objective in Eq. (17), we can plug in the optimal $\boldsymbol{\mu}_q$ and then solve the following simplified objective for $\boldsymbol{\Sigma}_q$:

$$J(q) = \frac{1}{2}\left(-\log|\boldsymbol{\Sigma}_q| - \text{const} + \sum_j \left[\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\right]_{jj}\right) \tag{18}$$

Suppose we chose $\boldsymbol{\Sigma}_q = \operatorname{diag}(\sigma_i^2)$. Then we get

$$J(q) = \frac{1}{2}\left(-\sum_j \log\sigma_j^2 - \text{const} + \sum_j \left[\sigma_j^2 \left(\boldsymbol{\Sigma}_p^{-1}\right)_{jj}\right]\right) \tag{19}$$

Setting $\frac{\partial J(q)}{\partial \sigma_i} = 0$, we get

$$\frac{\partial J(q)}{\partial \sigma_i} = -\frac{1}{\sigma_i} + \sigma_i \left(\boldsymbol{\Sigma}_p^{-1}\right)_{ii} = 0 \tag{20}$$

which gives

$$\sigma_i^{-2} = \left(\boldsymbol{\Sigma}_p^{-1}\right)_{ii} \tag{21}$$

which says that we should match the marginal precisions of the two distributions.

Applying Eq. (21) to Eq. (15) above, then:

$$\left(\sigma_{t|t}^i\right)^{-2} = \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^{\intercal}\mathbf{R}_t^{-1}\mathbf{H}_t\right)_{ii} \tag{22}$$

$$= \left(\sigma_{t|t-1}^i\right)^{-2} + \left(\mathbf{H}_t^{\intercal}\mathbf{R}_t^{-1}\mathbf{H}_t\right)_{ii} \tag{23}$$

and therefore the posterior covariance has the following form, matching Eq. (9):

$$\boldsymbol{\Sigma}_{t|t} = \operatorname{diag}\left(\left(\sigma_{t|t}^i\right)^2\right) = \operatorname{diag}\left(\left(\left(\sigma_{t|t-1}^i\right)^{-2} + \left(\mathbf{H}_t^{\intercal}\mathbf{R}_t^{-1}\mathbf{H}_t\right)_{ii}\right)^{-1}\right) \tag{24}$$

A.1.2. Mean

The mean update corresponds to the following equation:

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t \left(\mathbf{y}_t - \mathbf{h}(\boldsymbol{\mu}_{t|t-1})\right) \tag{25}$$

First, we can rewrite the Kalman gain from Eq. (6):

$$\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\intercal \left(\mathbf{R}_t + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\intercal\right)^{-1} \tag{26}$$

$$= \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\intercal\mathbf{R}_t^{-1} \left(\mathbf{I} + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_{\boldsymbol{\mu}_{t|t-1}}^\intercal\mathbf{R}_t^{-1}\right)^{-1} \tag{27}$$

We can use the **push-through identity**, which states the following:

$$U(I + VU)^{-1} = (I + UV)^{-1}U \tag{28}$$

Letting $U = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\intercal\mathbf{R}_t^{-1}$ and $V = \mathbf{H}_t$ and applying the identity to Eq. (27):

$$\mathbf{K}_t = \left(\mathbf{I} + \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_{\boldsymbol{\mu}_{t|t-1}}^\intercal\mathbf{R}_t^{-1}\mathbf{H}_t\right)^{-1} \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\intercal\mathbf{R}_t^{-1} \tag{29}$$

$$= \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_{\boldsymbol{\mu}_{t|t-1}}^\intercal\mathbf{R}_t^{-1}\mathbf{H}_t\right)^{-1} \mathbf{H}_t^\intercal\mathbf{R}_t^{-1} \tag{30}$$

$$= \boldsymbol{\Sigma}_{t|t}\mathbf{H}_t^\intercal\mathbf{R}_t^{-1} \tag{31}$$

$$\approx \operatorname{diag}\left(\left(\sigma_{t|t}^i\right)^2\right) \mathbf{H}_t^\intercal\mathbf{R}_t^{-1} \tag{32}$$

Therefore, the mean update can be re-expressed in the following form:

$$\boldsymbol{\mu}_{t|t} \approx \boldsymbol{\mu}_{t|t-1} + \operatorname{diag}\left(\left(\sigma_{t|t}^i\right)^2\right) \mathbf{H}_t^\intercal\mathbf{R}_t^{-1} \left(\mathbf{y}_t - \mathbf{h}(\boldsymbol{\mu}_{t|t-1})\right) \tag{33}$$

A.1.3. Complexity Analysis

Note that the bottleneck at each step is the computation of the diagonal matrix $\left(\mathbf{H}_t^\intercal\mathbf{R}_t^{-1}\mathbf{H}_t\right)_{ii}$ in the update step.

If we naively compute the full matrix product before taking the diagonal elements, it takes $O(N_z^2N_y + N_zN_y^2 + N_y^3)$ per step, where $\boldsymbol{\theta}_t \in \mathbb{R}^{N_z}$ and $\mathbf{y}_t \in \mathbb{R}^{N_y}$.

If $N_z \gg N_y$, as is usually the case when training deep neural networks, we can further optimize this using the following lemma.

Lemma 1. *Let $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times n}$. Then computing the diagonal elements of the product, $(AB)_{ii}$, takes $O(nm)$.*

Proof Note that we do not need to evaluate the full matrix product to take the diagonal elements. Computing each diagonal element takes time $O(m)$:

$$(AB)_{ii} = \sum_{j=1}^{m} A_{ij}B_{ji} \tag{34}$$

Since there are n diagonal elements, the total operation takes $O(nm)$.

Computation of $\mathbf{H}_t^\intercal\mathbf{R}_t^{-1}$ takes $O(N_zN_y^2 + N_y^3)$. Then since by Lemma 1, computing the diagonal elements of the product $(\mathbf{H}_t^\intercal\mathbf{R}_t^{-1})\mathbf{H}_t$ takes $O(N_zN_y)$, overall the computations take $O(N_zN_y^2)$.

A.2. Fully decoupled EKF

We now consider the inclusive KL objective: $D_{\mathbb{KL}}(p \parallel q)$:

$$D_{\mathbb{KL}}(p \parallel q) = \frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - D + \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right)^\intercal \boldsymbol{\Sigma}_q \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right) + \operatorname{tr}\left\{\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p\right\}\right) \tag{35}$$

The optimal solution for the mean is to set $\hat{\boldsymbol{\mu}}_q = \boldsymbol{\mu}_p$, which gives us the objective:

$$J(q) = \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_q| + \sum_j \left[\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p\right]_{jj}\right) + \text{const} \tag{36}$$

Choosing $\boldsymbol{\Sigma}_q = \operatorname{diag}(\sigma_i^2)$:

$$J(q) = \frac{1}{2}\left(\sum_j \log\sigma_j^2 + \sum_j \sigma_j^{-2}\left(\boldsymbol{\Sigma}_p\right)_{jj}\right) + \text{const} \tag{37}$$

Setting $\frac{\partial J(q)}{\partial \sigma_i} = 0$:

$$\frac{\partial J(q)}{\partial \sigma_i} = \frac{1}{\sigma_i} - \sigma_i^{-3}\left(\boldsymbol{\Sigma}_p\right)_{ii} = 0 \tag{38}$$

And therefore the solution is simply taking the diagonal elements of $\boldsymbol{\Sigma}_p$:

$$\sigma_i^2 = \left(\boldsymbol{\Sigma}_p\right)_{ii} \tag{39}$$

Plugging in the full-covariance EKF posterior covariance $\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_{t|t}$:

$$\begin{aligned}
\sigma_i^2 &= \left(\boldsymbol{\Sigma}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\right)_{ii} && (40)\\
&= \left(\boldsymbol{\Sigma}_{t|t-1}\right)_{ii} - \left(\mathbf{K}_t\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\right)_{ii} && (41)
\end{aligned}$$

Let $\mathbf{A}_i$ and $\mathbf{A}^i$ represent the ith row and column of matrix $\mathbf{A}$, respectively. Then, constraining the prior to be diagonal $\boldsymbol{\Sigma}_{t|t-1} = \operatorname{diag}(\sigma_i'^2)$:

$$\begin{aligned}
\sigma_i^2 &= \sigma_i'^2 - \left(\mathbf{K}_t\mathbf{H}_t\operatorname{diag}(\sigma_j'^2)\right)_{ii} && (42)\\
&= \sigma_i'^2 - \left(\mathbf{K}_t\mathbf{H}_t\right)_{ii}\sigma_i'^2 && (43)\\
&= \sigma_i'^2 - \left(\mathbf{K}_t\right)_i\mathbf{H}_t^i\sigma_i'^2 && (44)
\end{aligned}$$

Plugging in the full expression for $\mathbf{K}_t$:

$$\begin{aligned}
\left(\mathbf{K}_t\right)_i &= \left(\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\intercal\mathbf{S}^{-1}\right)_i = \left(\operatorname{diag}(\sigma_j'^2)\mathbf{H}_t^\intercal\mathbf{S}^{-1}\right)_i && (45)\\
&= \sigma_i'^2\left(\mathbf{H}_t^\intercal\mathbf{S}^{-1}\right)_i && (46)\\
&= \sigma_i'^2\left(\mathbf{H}_t^\intercal\right)_i\mathbf{S}^{-1} && (47)\\
&= \sigma_i'^2\left(\mathbf{H}_t^i\right)^\intercal\mathbf{S}^{-1} && (48)
\end{aligned}$$

In addition, we see that:

$$\mathbf{S} = \mathbf{R}_t + \mathbf{H}_t \mathrm{diag}(\sigma_j'^2)\mathbf{H}_t^\intercal = \mathbf{R}_t + \sum_j \mathbf{H}_t^j \sigma_j'^2 \left(\mathbf{H}_t^j\right)^\intercal \tag{49}$$

Plugging these in, therefore, we recover the covariance update equation for the fully-decoupled EKF, matching Eq. (11):

$$\sigma_i^2 = \sigma_i'^2 - \sigma_i'^2 \left(\mathbf{H}_t^i\right)^\intercal \left(\mathbf{R}_t + \sum_j \mathbf{H}_t^j \sigma_j'^2 \left(\mathbf{H}_t^j\right)^\intercal\right)^{-1} \mathbf{H}_t^i \sigma_i'^2 \tag{50}$$