Appendix to: Fast Kalman filtering and forward-backward smoothing via a low-rank perturbative approach

Eftychios A. Pnevmatikakis, Kamiar Rahnama Rad, Jonathan Huggins and Liam Paninski

October 15, 2012

A Extension to nonlinear observations

We would like to incorporate observations y_t obeying an arbitrary conditional density $p(y_t|x_t)$ into our filter equations. This is difficult in general, since if $p(y_t|x_t)$ is chosen maliciously the posterior $p(x_t|Y_{1:t})$ may be highly non-Gaussian, and our basic Kalman recursion will break down. However, if $\log p(y_t|x_t)$ is a smooth, concave function of x_t, it is known that a Gaussian approximation to $p(x_t|Y_{1:t})$ will often be fairly accurate (Fahrmeir and Tutz, 1994; Brown et al., 1998; Paninski et al., 2010), and our recursion may be adapted in a fairly straightforward manner.

For simplicity, we will focus on the case that the observations y_{it} are independent samples from $p(y_{it}|[B_t]_i x_t)$, where $[B_t]_i$ denotes the i-th row of the observation matrix B_t. (The extension to the case that y_t depends in a more general manner on the projection $B_t x_t$ may be handled similarly.) We approximate the posterior mean μ_t with the one-step maximum a posteriori (MAP) estimate,

$$\begin{aligned}\mu_t &\approx \arg\max_{x_t} \left[\log p(x_t|Y_{1:t-1}) + \log p(y_t|x_t)\right] \\ &= \arg\max_{x_t} \left[-\frac{1}{2}(x_t - m_t)^T \tilde{P}_t^{-1}(x_t - m_t) + \sum_i \log p(y_{it}|[B_t]_i x_t)\right].\end{aligned} \tag{S-1}$$

(Recall that the one-step predictive covariance matrix $\tilde{P}_t$ and mean m_t were defined in (4) and (5) in the main text. This MAP update is exact in the linear-Gaussian case (and corresponds to the Kalman filter), but is an approximation more generally. To compute this MAP estimate, we use Newton's method. We need the gradient and Hessian of the log-posterior with respect to x_t,

$$\nabla_t = -\tilde{P}_t^{-1}(x_t - m_t) + B_t^T f_1(x_t)$$
$$[H]_{tt} = -\tilde{P}_t^{-1} + B_t^T \mathrm{diag}\{f_2(x_t)\} B_t,$$

respectively. Here $f_1(x_t)$ and $f_2(x_t)$ are the vectors formed by taking the first and second derivatives, respectively, of $\log p(y_{it}|u)$ at $u = B_i x_t$, with respect to u. Now we may form the Newton step:

$$\begin{aligned}
x_{new} =& x_{old} - \left(-\tilde{P}_t^{-1} + B_t^T \mathrm{diag}\{f_2(x_{old})\} B_t\right)^{-1} \left(-\tilde{P}_t^{-1}(x_{old} - m_t) + B_t^T f_1(x_{old})\right) \\
=& x_{old} - (\tilde{P}_t - \tilde{P}_t B_t^T(-\mathrm{diag}\{f_2(x_{old})^{-1}\} + B_t \tilde{P}_t B_t^T)^{-1} B_t \tilde{P}_t) \left[\tilde{P}_t^{-1}(x_{old} - m_t) - B_t^T f_1(x_{old})\right] \\
=& m_t + \tilde{P}_t B_t^T(-\mathrm{diag}\{f_2(x_{old})^{-1}\} + B_t \tilde{P}_t B_t^T)^{-1} B_t \left[x_{old} - m_t - \tilde{P}_t B_t^T f_1(x_{old})\right] + \tilde{P}_t B_t^T f_1(x_{old})
\end{aligned}$$

We iterate, using a backstepping linesearch to guarantee that the log-posterior increases on each iteration, until convergence (i.e., when $x_{new} \approx x_{old}$, set $\mu_t = x_{new}$). Then, finally, we update the covariance C_t by replacing W_t^{-1} with $-\mathrm{diag}\{f_2(x_t)\}$ in the original derivation. Since multiplication by $\tilde{P}_t$ is assumed fast (and we need to compute $\tilde{P}_t B_t^T$ just once per timestep), all of these computations remain tractable.

A similar methodology can also be derived for the case where we are interested in the full forward-backward smoothing. It is not hard to modify Theorem 4.1 for the case of non-linear observations. As a result, an iterative search direction algorithm can also be applied in this setup to find the MAP estimate. The algorithm corresponds now to an inexact Newton's method (Dembo et al., 1982; Sun and Yuan, 2006) (as opposed to steepest descent in the quadratic case). Since by controlling the threshold we can make the Hessian approximation error arbitrarily small, the algorithm is guaranteed to converge (Eisenstat and Walker, 1994; Sun and Yuan, 2006). Some details can be found in Pnevmatikakis and Paninski (2012).

Finally, we note that it is also straightforward to adapt these fast methods for sampling from the posterior $p(X|Y)$ once the MAP path for X is obtained. This can be done either in the context of the filter-forward sample-backward approach discussed in Jungbacker and Koopman (2007) or via the perturbed-MAP sampling approach discussed in Papandreou and Yuille (2010); however, we have not yet pursued this direction extensively.

B Effective rank

Here we take a closer look at the notion of the effective rank, which characterizes the scaling properties of our algorithm. We examine the effective rank of the matrices $Z_t^{-1/2} U_t C_{0,t}$, where Z_t is defined in (18) of the main text, and also derive heuristic methods that lead to tighter bounds for the effective rank. Finally, we present an example that supports our arguments. Although the analysis here focuses on the fast KF algorithm, similar results hold for the LRBT algorithm as well.

We will use the following approximation that can be derived by Taylor-expanding the inverse of a matrix. For a scalar ε with $|\varepsilon| \ll \|A\|, \|B\|$, and A, $A + \varepsilon B$ invertible matrices, it holds that

$$(A + \varepsilon B)^{-1} = A^{-1} - \varepsilon A^{-1} B A^{-1} + O(\varepsilon^2), \tag{S-2}$$

B.1 Proof of Proposition 3.2

Proof. The proof uses induction and the Woodbury lemma. The statement is trivial for $t = 1$. Suppose that (17) holds for $t = k$. Then by using the abbreviation $\Omega_k = U_k C_{0,k} A^T$, and applying the Woodbury lemma we have for $t = k + 1$

$$
\begin{aligned}
C_{k+1}^{-1} =&(AC_kA^T+V)^{-1}+B_{k+1}^TW^{-1}B_{k+1}\\
=&(A(C_{0,k}^{-1}+U_k^TF_kU_k)^{-1}A^T+V)^{-1}+B_{k+1}^TW^{-1}B_{k+1}\\
\overset{(w)}{=}&(A(C_{0,k}-C_{0,k}U_k^T(F_k^{-1}+U_kC_{0,k}U_k^T)^{-1}U_kC_{0,k})A^T+V)^{-1}+B_{k+1}^TW^{-1}B_{k+1}\\
=&(C_{0,k+1}-\Omega_k^T(F_k^{-1}+U_kC_{0,k}U_k^T)^{-1}\Omega_k)^{-1}+B_{k+1}^TW^{-1}B_{k+1}\\
\overset{(w)}{=}&C_{0,k+1}^{-1}+B_{k+1}^TW^{-1}B_{k+1}+C_{0,k+1}^{-1}\Omega_k^T(F_k^{-1}+U_kC_{0,k}U_k^T-\Omega_kC_{0,k+1}^{-1}\Omega_k^T)^{-1}\Omega_kC_{0,k+1}^{-1}\\
=&C_{0,k+1}^{-1}+U_{k+1}^TF_{k+1}U_{k+1},
\end{aligned}
$$

where $\overset{(w)}{=}$ indicate applications of the Woodbury lemma. The proposition follows by taking the inverse and applying the Woodbury lemma once more. □

B.2 Effective rank of $Z_t^{-1/2}U_tC_{0,t}$

To develop an analytically tractable example, we assume that the measurement matrices B_t are $b\times d$ i.i.d. random matrices, where each entry is symmetric with zero mean and variance $1/d$. For simplicity, we assume that the observation noise covariance matrix is the same at all times ($W_t = W$). The matrix Z_t ((18) in the main text) can then be written as

$$
Z_t = I_t\otimes W + U_tC_{0,t}U_t^T + \mathrm{blkdiag}\{\mathbf{0}_b, U_{t-1}C_{0,t-1}U_{t-1}^T\} + \ldots + \mathrm{blkdiag}\{\mathbf{0}_b,\ldots,\mathbf{0}_b, U_1C_{0,1}U_1^T\},
$$

where I_t is a $t\times t$ identity matrix, $\mathbf{0}_b$ is a $b\times b$ all zero matrix, and $\otimes$ denotes the Kronecker product. Assuming that $C_{0,t}=C_0$ for all t and using (21), we can write Z_t as

$$
Z_t = I_t\otimes W + \underbrace{\mathrm{blkdiag}\left\{B_tC_0B_t^T, B_{t-1}(C_0+A^TC_0A)B_{t-1}^T,\ldots,B_1\left(\sum_{i=0}^{t-1}(A^T)^iC_0A^i\right)B_1^T\right\}}_{J} + K, \tag{S-3}
$$

where K is the matrix that includes the non-diagonal blocks of the products $U_l C_0 U_l^T$. For $m = n$ we have $[K]_{mn} = 0$, while for $m \neq n$, the mn-th block of K is given by

$$[K]_{mn} = \sum_{l=1}^{\min(m,n)} B_{t+1-m} (A^T)^{m-l} C_0 A^{n-l} B_{t+1-n}^T. \tag{S-4}$$

Proceeding as before, we can bound the effective rank by the minimum number of blocks required to capture a θ fraction of $\mathbb{E}\|Z_t^{-1/2} U_t C_{0,t}\|_F^2$. To compute this expected energy, we first argue that with high probability J is much larger than K (in a suitable sense) for large d. To see why, assume at first for simplicity that $b = 1$, and that the matrices A and C_0 are proportional to the identity. Then a quick calculation shows that each block of J will be composed of weighted chi-squared variables with d degrees of freedom; thus the means of these variables will be bounded away from zero, while the variance decreases linearly as a function of $1/d$. On the other hand, each block of K will be a zero mean random variable (since B_t and B_s are zero-mean and independent for $s \neq t$), with variance decreasing linearly with $1/d$. In addition, the variance of the elements of K decreases exponentially away from the diagonal $m = n$, due to the effect of the repeated multiplication by A in (S-4); thus K is effectively banded, with bandwidth determined by the largest singular value of A. The effectively banded nature of K implies that $J \gg K$ in terms of suitable matrix norms, for sufficiently large d. The same argument applies in the general case, where each block of J will be distributed according to a Wishart distribution (note that every term $(A^T)^m C_0 A^m$ is a PD matrix) and thus its expected value is bounded away from zero while its covariance matrix tends to zero as d increases, whereas each block of K will be a zero mean random variable with variance decreasing in d. We revisit this approximation in the example in section B.4.

By denoting $\Xi = I_t \otimes W + J$, the expected energy $\mathbb{E}\|Z_t^{-1/2} U_t C_{0,t}\|_F^2$ can then be approximated as

$$\mathbb{E}\|Z_t^{-1/2} U_t C_{0,t}\|_F^2 \approx \mathbb{E}\|\Xi^{-1/2} U_t C_0\|_F^2. \tag{S-5}$$

The expected energy $\mathbb{E}\|Z_t^{-1/2} U_t C_{0,t}\|_F^2$ cannot be computed in closed form. However, we conjecture that the number of blocks we need to keep from $Z_t^{-1/2} U_t C_{0,t}$ to capture a θ fraction of its energy is bounded from above by the number of blocks we need to keep from $U_t C_{0,t}$. To get some intuition about this observe the structure of J in (S-3). J (and therefore Ξ) is a block diagonal matrix where the expected energy of each block increases. Consequently Ξ^{-1}, and therefore $Z_t^{-1/2}$ can be approximated by a block diagonal matrix where the energy of each block decreases. As a result, when multiplying $U_t C_0$ with $Z_t^{-1/2}$, the energy of the blocks of $Z_t^{-1/2} U_t C_{0,t}$ would decrease faster than the energy of the blocks of $U_t C_{0,t}$ and hence fewer blocks would be required to capture a certain fraction. To justify our conjecture, note that Ξ is a block diagonal matrix, and therefore by using (S-5) we approximate

$$\mathbb{E}\|[Z_t^{-1/2} U_t C_{0,t}]_{m+1}\|_F^2 \approx \mathbb{E}\left\| \left(W + B \left(\sum_{i=0}^{m} (A^T)^i C_0 A^i \right) B^T \right)^{-1/2} B (A^T)^m C_0 \right\|_F^2 . \tag{S-6}$$

Our conjecture will hold if the expected energy of the blocks of $Z_t^{-1/2} U_t C_{0,t}$ drops faster than in the blocks of $U_t C_{0,t}$, since then the energy of $Z_t^{-1/2} U_t C_{0,t}$ will be concentrated in fewer blocks and therefore fewer singular values will be required to capture a certain fraction of this energy. In mathematical terms, this can be expressed as

$$\frac{\mathbb{E}\|[U_t C_{0,t}]_{m+1}\|_F^2}{\mathbb{E}\|[U_t C_{0,t}]_m\|_F^2} \geq \frac{\mathbb{E}\|[Z_t^{-1/2} U_t C_{0,t}]_{m+1}\|_F^2}{\mathbb{E}\|[Z_t^{-1/2} U_t C_{0,t}]_m\|_F^2} \tag{S-7}$$

or using (23) and (S-6)

$$\frac{\mathbb{E}\|B(A^T)^m C_0\|_F^2}{\mathbb{E}\|B(A^T)^{m-1} C_0\|_F^2} \geq \frac{\mathbb{E}\left\| \left(W + B \left(\sum_{i=0}^{m} (A^T)^i C_0 A^i \right) B^T \right)^{-1/2} B (A^T)^m C_0 \right\|_F^2}{\mathbb{E}\left\| \left(W + B \left(\sum_{i=0}^{m-1} (A^T)^i C_0 A^i \right) B^T \right)^{-1/2} B (A^T)^{m-1} C_0 \right\|_F^2} . \tag{S-8}$$

We can now continue our analysis for two different cases. First we show that (S-8) holds if we are in the low signal-to-noise ratio (SNR) regime. We then show that (S-8) holds and provide a bound for the effective rank of $Z_t^{-1/2} U_t C_{0,t}$ in the case where we take only one measurement per timestep ($b = 1$), and also A, V are proportional to the identity matrix.

B.2.1 The low-SNR case

We can consider that we are in the low SNR regime if $\|W\| \gg \|V\|, \|A\|, \|B^T C_0 B\|$, i.e., the measurement noise covariance matrix is much larger than that of the state variable and the projection of the state variable onto the measurement matrices. We assume that the noise of the observations is i.i.d. with variance σ^2, i.e., $W = \sigma^2 I$. By denoting by V^m the matrices $\sum_{i=0}^{m}(A^T)^i C_0 A^i$ we have that

$$\begin{aligned}
\mathbb{E}\|(W + BV^mB^T)^{-1/2}B(A^T)^mC_0\|_F^2 &= \mathrm{Tr}\mathbb{E}\left(C_0A^mB^T(W + BV^mB^T)^{-1}B(A^T)^mC_0\right) \\
&\overset{\text{(S-2)}}{\approx} \mathrm{Tr}\mathbb{E}\left(C_0A^mB^TW^{-1}B(A^T)^mC_0\right) \\
&\quad - \underbrace{\mathrm{Tr}\mathbb{E}\left(C_0A^mB^TW^{-1}BV^mB^TW^{-1}B(A^T)^mC_0\right)}_{f_m} \\
&= \mathbb{E}\|W^{-1/2}B(A^T)^mC_0\|_F^2 - f_m,
\end{aligned} \tag{S-9}$$

After some algebra we have that

$$\begin{aligned}
f_m &= \frac{b}{d^2\sigma^4}\mathrm{Tr}\left(C_0A^m((b+1)V^m + \mathrm{Tr}(V^m)I)(A^T)^mC_0\right). \\
&= \frac{b}{d^2\sigma^4}\left(\underbrace{(b+1)\sum_{i=1}^{d}\left(c_i^2\alpha_i^{2m}\left(\sum_{l=0}^{m}c_i\alpha_i^{2l}\right)\right)}_{f_m^1} + \underbrace{\left(\sum_{i=1}^{d}\sum_{l=0}^{m}c_i\alpha_i^{2l}\right)\sum_{i=1}^{d}c_i^2\alpha_i^{2m}}_{f_m^2}\right).
\end{aligned} \tag{S-10}$$

Plugging (S-9) into (S-8), we see that (S-8) is equivalent to

$$\frac{\mathbb{E}\|B(A^T)^mC_0\|_F^2}{\mathbb{E}\|B(A^T)^{m-1}C_0\|_F^2} \le \frac{f_m}{f_{m-1}}. \tag{S-11}$$

To show (S-11) it is sufficient to show that

$$\frac{\mathbb{E}\|B(A^T)^mC_0\|_F^2}{\mathbb{E}\|B(A^T)^{m-1}C_0\|_F^2} \le \frac{f_m^1}{f_{m-1}^1} \quad \text{and} \quad \frac{\mathbb{E}\|B(A^T)^mC_0\|_F^2}{\mathbb{E}\|B(A^T)^{m-1}C_0\|_F^2} \le \frac{f_m^2}{f_{m-1}^2}. \tag{S-12}$$

Both of these inequalities are easy to show:

$$\frac{f_m^1}{f_{m-1}^1} = \frac{\sum_{i=1}^d c_i^2 \alpha_i^{2m} \left(\sum_{l=0}^m c_i \alpha_i^{2l}\right)}{\sum_{i=1}^d c_i^2 \alpha_i^{2(m-1)} \left(\sum_{l=0}^{m-1} c_i \alpha_i^{2l}\right)} \geq \frac{\sum_{i=1}^d c_i^2 \alpha_i^{2m}}{\sum_{i=1}^d c_i^2 \alpha_i^{2(m-1)}} = \frac{\mathbb{E}\|B(A^T)^m C_0\|_F^2}{\mathbb{E}\|B(A^T)^{m-1} C_0\|_F^2}$$
$$\frac{f_m^2}{f_{m-1}^2} = \frac{\left(\sum_{i=1}^d \sum_{l=0}^m c_i \alpha_i^{2l}\right) \sum_{i=1}^d c_i^2 \alpha_i^{2m}}{\left(\sum_{i=1}^d \sum_{l=0}^{m-1} c_i \alpha_i^{2l}\right) \sum_{i=1}^d c_i^2 \alpha_i^{2(m-1)}} \geq \frac{\sum_{i=1}^d c_i^2 \alpha_i^{2m}}{\sum_{i=1}^d c_i^2 \alpha_i^{2(m-1)}} = \frac{\mathbb{E}\|B(A^T)^m C_0\|_F^2}{\mathbb{E}\|B(A^T)^{m-1} C_0\|_F^2}. \tag{S-13}$$

Therefore (S-12) holds which implies (S-11) and our conjecture (S-8). Note in the case of very low SNR ($\sigma^2 \to \infty$), (S-8) becomes an equality since in the right side we have $W + B\left(\sum_{i=0}^m (A^T)^i C_0 A^i\right) B^T \approx \sigma^2 I$.

B.2.2 The case of $b = 1$ and $A, V \propto I_d$

As a second case, we can analyze the case where we have only one observation per timestep and the matrices A, V are proportional to the identity. In this case, we can write $A = \alpha I$, $V = vI$ and $W = \sigma^2$, and we also have $C_0 = cI$ with $c = v(1-\alpha^2)^{-1}$. Using the fact that A, V and C_0 are proportional to the identity, the expected energy of the $(m+1)$-th block of $Z_t^{-1/2} U_t C_{0,t}$ can then be written as

$$\begin{aligned}
\mathbb{E}\left\|[Z_t^{-1/2} U_t C_0]_{m+1}\right\|_F^2 &\overset{\text{(S-6)}}{\approx} \mathbb{E}\left\| \left(W + B\left(\sum_{i=0}^m (A^T)^i C_0 A^i\right) B^T\right)^{-1/2} B(A^T)^m C_0 \right\|_F^2 \\
&= \mathrm{Tr}\mathbb{E}\left(c\alpha^m B^T (\sigma^2 + c\sum_{i=0}^m \alpha^{2i} BB^T)^{-1} Bc\alpha^m \right) \\
&= \frac{c\alpha^{2m}}{\sum_{i=0}^m \alpha^{2i}} \mathbb{E}\left(\frac{BB^T}{\frac{\sigma^2}{c\sum_{i=0}^m \alpha^{2i}} + BB^T} \right)
\end{aligned} \tag{S-14}$$

Since $b = 1$, BB^T is a χ^2-random variable with d degrees of freedom. Using Mathematica (Wolfram Research, Inc., 2010), the above expectation can be computed in closed form and we have

$$\mathbb{E}\left\|[Z_t^{-1/2} U_t C_0]_{m+1}\right\|_F^2 \approx \frac{c\alpha^{2m}}{2\sum_{i=0}^m \alpha^{2i}} d \exp\left(\frac{\sigma^2}{2c\sum_{i=0}^m \alpha^{2i}}\right) \mathrm{Ei}_{1+d/2}\left(\frac{\sigma^2}{2c\sum_{i=0}^m \alpha^{2i}}\right), \tag{S-15}$$

where $\mathrm{Ei}_n(\cdot)$ denotes the generalized exponential integral function of n-th order. For $x \geq 0, n \geq 1$, $\mathrm{Ei}_n(x)$ can be tightly bound as (Abramowitz and Stegun, 1964)

$$\frac{1}{x+n} < e^x \mathrm{Ei}_n(x) < \frac{1}{x+n-1}. \tag{S-16}$$

Using this bound, $\mathbb{E}\left\|[Z_t^{-1/2} U_t C_0]_{m+1}\right\|_F^2$ can be bounded as

$$\frac{dc^2\alpha^{2m}}{\sigma^2 + (d+2)c\sum_{i=0}^{m}\alpha^{2i}} < \mathbb{E}\left\|[Z_t^{-1/2} U_t C_0]_{m+1}\right\|_F^2 < \frac{dc^2\alpha^{2m}}{\sigma^2 + dc\sum_{i=0}^{m}\alpha^{2i}} \tag{S-17}$$

To compute a bound for $z_\theta(Z_t^{-1/2} U_t C_0)$ we need to find the minimum integer k such that

$$\sum_{l=1}^{k} \mathbb{E}\left\|[Z_t^{-1/2} U_t C_0]_l\right\|_F^2 \geq \theta \sum_{l=1}^{\infty} \mathbb{E}\left\|[Z_t^{-1/2} U_t C_0]_l\right\|_F^2 \overset{\text{(S-17)}}{\Rightarrow} \sum_{l=0}^{k-1} \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}} \geq \theta \sum_{l=0}^{\infty} \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}} \tag{S-18}$$

with

$$v' = 1 + \sigma^2 \frac{1-\alpha^2}{dc}. \tag{S-19}$$

The sums in (S-18) cannot be computed in closed form. However, we can approximate them with integrals:

$$\min\left\{k : \sum_{l=0}^{k-1} \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}} \geq \theta \sum_{l=0}^{\infty} \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}}\right\} \approx \min\left\{k : \int_0^k \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}}\, dl \geq \theta \int_0^\infty \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}}\, dl\right\} \tag{S-20}$$

Using the formula

$$\int_{l=0}^{k} \frac{\alpha^{2l}}{v' - \alpha^{2(l+1)}}\, dl = \frac{\log(v' - \alpha^2) - \log(v' - \alpha^{2(k+1)})}{\alpha^2 \log(\alpha^2)}, \tag{S-21}$$

we can calculate the required number of blocks as

$$k_Z \approx \left\lceil \frac{\log\left(v' - (v' - \alpha^2)\left(1 - \alpha^2/v'\right)^{-\theta}\right)}{2\log(|\alpha|)} - 1 \right\rceil. \tag{S-22}$$

A simple algebraic calculation shows that (S-22) is increasing with σ^2 and we also have that

$$\lim_{\sigma^2\to\infty} k_Z = \left\lceil \frac{\log(1-\theta)}{2\log(|\alpha|)} \right\rceil, \tag{S-23}$$

i.e. in the limiting case where $\sigma^2 \to \infty$ where $z_\theta(Z_t^{-1/2}U_tC_{0,t}) \to z_\theta(U_tC_{0,t})$, our approximate bound agrees with the computed bound for k_U in the main text.

B.3 Heuristic bounds for the effective rank of $G_t^{1/2}$

As explained before, the computation of the effective rank of $G_t^{1/2}$ is hard, since G_t is obtained from a series of successive low-rank approximations. For very small t, when just a few measurements are available, we expect that the effective rank grows as $z_\theta(G_t^{1/2}) \approx bt$, since no measurement is "old enough to be forgotten." However, as t grows, the effective rank saturates and concentrates around a specific value. When that happens, at each step a number of b new measurements are taken and also a similar number of b linear combinations of old measurements is dropped since they contribute very little. In that case, if the effective rank saturates at a value bk, then G_t is approximately of size $b(k+1)$, and bk singular values suffice to capture a θ-fraction of its energy. Since the matrices G_t are hard to work with we can apply this argument to the matrices $U_tC_{0,t}$ and $Z_t^{-1/2}U_tC_{0,t}$ to derive two heuristic bounds for $z_\theta(G_t^{1/2})$. For the first bound, we will look for the minimum k such that the first k blocks of $U_tC_{0,t}$ capture a θ fraction of the expected energy of the first $k+1$ blocks of $U_tC_{0,t}$. Hence, instead of solving (22) of the main text we seek the solution to

$$k_{G^1} = \min_{l\in\mathbb{N}}\{l : \mathbb{E}\|[U_t]_{1:l}C_{0,t}\|_F^2 \geq \theta\mathbb{E}\|[U_t]_{1:l+1}C_{0,t}\|_F^2\}. \tag{S-24}$$

The solution of (S-24) approximates this saturation point by finding the minimum integer k_{G^1} such that the expected energy of the first k_{G^1} blocks of $U_tC_{0,t}$ capture a θ fraction of the expected energy of the first $k_{G^1}+1$ blocks of $U_tC_{0,t}$. Solving (S-24) in a similar way and assuming $b=1$ gives

$$k_{G^1} = \left\lceil \frac{\log(1-\theta) - \log(1-\alpha^2\theta)}{2\log(\alpha)} \right\rceil. \tag{S-25}$$

For the second heuristic bound, we can similarly seek the solution to

$$k_{G^2} = \min_{l\in\mathbb{N}}\{l : \mathbb{E}\|[Z_t^{-1/2}U_t]_{1:l}C_{0,t}\|_F^2 \geq \theta\mathbb{E}\|[Z_t^{-1/2}U_t]_{1:l+1}C_{0,t}\|_F^2\}. \tag{S-26}$$

To find k_{G^2} we solve a modified version of (S-18)

$$k_{G^2} = \min_{l\in\mathbb{N}}\left\{l : \sum_{m=0}^{l-1}\frac{\alpha^{2m}}{v' - \alpha^{2(m+1)}} \geq \theta\sum_{m=0}^{l}\frac{\alpha^{2m}}{v' - \alpha^{2(m+1)}}\right\}. \tag{S-27}$$

Using similar approximations we have that

$$k_{G^2} = \min_{l\in\mathbb{N}}\left\{l : \theta \leq \frac{\log(v' - a^2) - \log(v' - a^{2(l+1)})}{\log(v' - a^2) - \log(v' - a^{2(l+2)})}\right\}. \tag{S-28}$$

Eq. (S-28) cannot be solved in closed form. However we note that k_{G^2} is increasing with σ^2, $k_{G^2} \leq k_{G^1}$ and in the limit case where $\sigma^2 \to \infty$, (S-28) converges to (S-25). Moreover, it is interesting to see that in the case when $\alpha \to 1$, the two bounds become equal and we have

$$\lim_{\alpha\to 1} k_{G^1} = \lim_{\alpha\to 1} k_{G^2} = \left\lceil\frac{\theta}{1-\theta}\right\rceil. \tag{S-29}$$

Although this result shows that the effective rank stays bounded, our algorithm is not applicable in the case where $\|A\| = 1$. Our approximation results come from the assumption that the information from measurements at time t decays exponentially as we move away from t. This assumption is valid only in the case when $\|A\| < 1$. We revisit this issue in appendix C (see remark C.3).

Another interesting way to derive the heuristic bounds is by performing a series of truncations on the matrices $U_tC_{0,t}$ or $Z_t^{-1/2}U_tC_{0,t}$ in the case of very large t ($t \to \infty$). For example k_{G^1} can be obtained if we consider the recursion $k_{G^1}^n$ defined as

$$k_{G^1}^{n+1} = \min_{l\in\mathbb{N}}\{l : \mathbb{E}\|[U_t]_{1:l}C_{0,t}\|_F^2 \geq \theta\mathbb{E}\|[U_t]_{1:k_{G^1}^n}C_{0,t}\|_F^2\} \tag{S-30}$$

with
$$k_{G^1}^1 = \min_{l\in\mathbb{N}}\{l : \mathbb{E}\|[U_t]_{1:l}C_{0,t}\|_F^2 \geq \theta\mathbb{E}\|U_tC_{0,t}\|_F^2\}. \tag{S-31}$$

In this case, we have

$$k_{G^1}^n \leq \left\lceil \frac{\log(1-\theta) + \log(1-(a^2\theta)^n) - \log(1-\alpha^2\theta)}{\log a^2} \right\rceil \stackrel{n\to\infty}{\longrightarrow} k_{G^1}. \tag{S-32}$$

Numerical simulations also establish a similar result for k_{G^2}.

B.4 An example

To illustrate our analysis we present a simple smoothing example in which we picked $T = 500, d = 440, b = 1, A = 0.95I_d, W = \sigma^2 = 0.5, V = 0.1I_d$. In Fig. S-1 we mark the effective rank of the matrix G_t^θ for $t = 495$ for 5 different values of the threshold value θ. The θ-superscript here denotes the threshold that was used in the LRBT recursion to derive the matrix G_t. In particular we mark the values $z_\theta((G_t^\theta)^{1/2})$, for five different values of θ, which correspond to the actual rank used in the LRBT algorithm. These values can be approximated by the heuristic bounds k_{G^1} and k_{G^2}, derived in (S-25) and (S-28) respectively (dashed-dotted lines). We also plot the effective rank of the matrices $U_t\tilde{D}_t^{-1}, Z_t^{-1/2}U_t\tilde{D}_t^{-1}$ (dashed lines) as well as their theoretical bounds k_U and k_Z of (24) in the main text and (S-22) respectively (solid lines).

Fig. S-1 shows that the theoretical bounds k_U and k_Z provide relatively tight upper bounds for the actual effective rank of $U_t\tilde{D}_t^{-1}$ and $Z_t^{-1/2}U_t\tilde{D}_t^{-1}$ respectively. In addition, the heuristic bounds of for $z_\theta((G_t^\theta)^{1/2})$, especially (S-28), provide a good approximation of the actual effective rank used in the algorithm. All the bounds describe worst case scenarios and become looser when A is not proportional to the identity, in which case some eigenvalues decay faster than others. We finally plot in logarithmic scale the magnitude of the entries of the matrix Z_t in Fig. S-1 (right). We see that most of the energy of Z_t (96.4%) is concentrated in the main diagonal. As a result, the approximation of (S-5) which is important for our analysis, is justified.

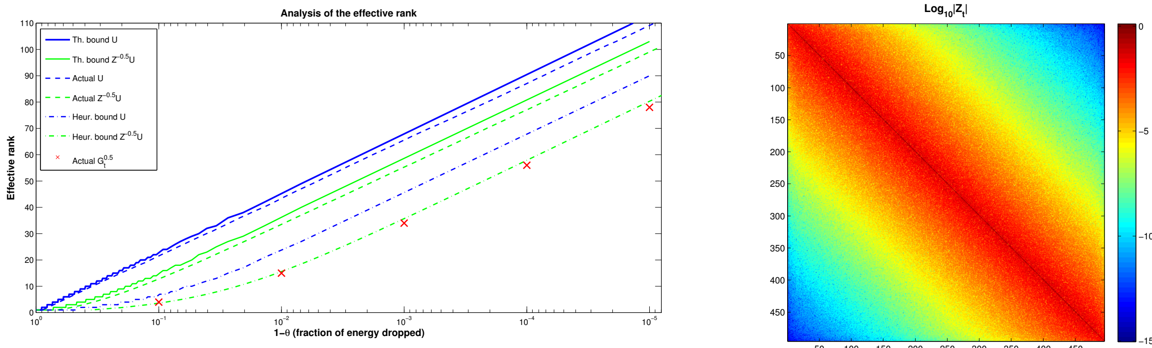

Figure S-1: Left: Analysis of the effective rank. Solid blue/green: Theoretical bounds on $z_\theta(U_tC_{0,t})$ (k_U (24) in the main text), and $z_\theta(Z_t^{-1/2}U_tC_{0,t})$ (k_Z eq. (S-22)). Dashed blue/green: Actual $z_\theta(U_tC_{0,t})$ and $z_\theta(Z_t^{-1/2}U_tC_{0,t})$, respectively. Dash-dotted blue/green. Heuristic bounds on $z_\theta(G_t^{1/2})$ based on eqs. (S-25) (k_{G^1}) and (S-28) (k_{G^2}) respectively. The marked points correspond to the actual effective rank of $G_t^{1/2}$, when the LRBT algorithm was run for five different threshold values. Note that the heuristic bound of (S-28) provides a very good approximation for the effective rank, and characterizes the computational gains of the algorithm. Right: Magnitude of the entries of Z_t in logarithmic scale. The energy of Z_t is concentrated in the main diagonal.

B.5 Effective rank for the LRBT algorithm

Finally, we examine the effective rank of the matrices involved in the LRBT algorithm. Using a similar induction method as in the fast KF case, the matrix M_t can be written as

$$M_t = \tilde{D}_t + U_t^T F_t U_t \Rightarrow M_t^{-1} = \tilde{D}_t^{-1} - \tilde{D}_t^{-1} U_t^T (F_t^{-1} + U_t \tilde{D}_t^{-1} U_t^T)^{-1} U_t \tilde{D}_t^{-1}, \tag{S-33}$$

where the matrices U_t and F_t are defined recursively (note again the recycled notation for U_t and F_t) as follows:

$$U_t = \begin{bmatrix} B_t \\ U_{t-1}\tilde{D}_{t-1}^{-1}E_{t-1} \end{bmatrix}, \quad F_t^{-1} = \begin{bmatrix} W_t & 0 \\ 0 & F_{t-1}^{-1} + U_{t-1}\tilde{D}_{t-1}^{-1}U_{t-1}^T \end{bmatrix}, \tag{S-34}$$

with $U_1 = B_1$ and $F_1 = W_1^{-1}$. It is easy to see that for large t, $\tilde{D}_t$ converges to V^{-1}. As a result the recursion of (S-34) can be approximated as

$$U_t \approx [B_t^T \quad AU_{t-1}^T]^T, \quad F_t^{-1} \approx \text{blkdiag}\{W, F_{t-1}^{-1} + U_{t-1}VU_{t-1}^T\}. \tag{S-35}$$

Note that (S-35) is identical to (21) in the main text. Moreover, the general form of M_t^{-1} (S-33) is the same as the general form C_t ((17) in the main text), with the only difference that $C_{0,t}$ has been replaced with $\tilde{D}_t^{-1}$, which does not affect the scaling properties of the effective rank. Therefore the effective rank for the LRBT case can estimated using the same analysis as in the fast KF filter case.

C Convergence properties of the LRBT algorithm

C.1 Proof of theorem 4.1

We can write the forward-backward recursion of the Block-Thomas algorithm in matrix-vector form. The backward recursion can be expressed as

$$\left.\begin{aligned} &\mathbf{s}_T = \boldsymbol{q}_T, \\ &\mathbf{s}_t = \boldsymbol{q}_t + \Gamma_t \mathbf{s}_{t+1}, \quad t = T-1, \dots, 1 \end{aligned}\right\} \Rightarrow \begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_{T-1} \\ \boldsymbol{s}_T \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & \Gamma_1 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \Gamma_{T-1} \\ 0 & 0 & \dots & 0 \end{bmatrix}}_{\Gamma} \begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_{T-1} \\ \boldsymbol{s}_T \end{bmatrix} + \begin{bmatrix} \boldsymbol{q}_1 \\ \vdots \\ \boldsymbol{q}_{T-1} \\ \boldsymbol{q}_T \end{bmatrix}. \tag{S-36}$$

Similarly, the forward recursion

$$\begin{aligned} \boldsymbol{q}_1 &= -M_1^{-1}\nabla_1, \\ \boldsymbol{q}_t &= -M_t^{-1}(\nabla_t - E_{t-1}^T \boldsymbol{q}_{t-1}), \ t = 2, \dots, T \end{aligned} \tag{S-37}$$

can be written in matrix-vector form as

$$\begin{bmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \vdots \\ \boldsymbol{q}_T \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & \dots & 0 & 0 \\ M_2^{-1}E_1^T & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & M_T^{-1}E_{T-1}^T & 0 \end{bmatrix}}_{E} \begin{bmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \vdots \\ \boldsymbol{q}_T \end{bmatrix} - \underbrace{\begin{bmatrix} M_1^{-1} & 0 & \dots & 0 \\ 0 & M_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & M_T^{-1} \end{bmatrix}}_{M^{-1}} \begin{bmatrix} \boldsymbol{\nabla}_1 \\ \boldsymbol{\nabla}_2 \\ \vdots \\ \boldsymbol{\nabla}_T \end{bmatrix} \tag{S-38}$$

Combining (S-36) and (S-38) we have

$$\mathbf{s} = -(I - \Gamma)^{-1}(I - E)^{-1}M^{-1}\nabla, \tag{S-39}$$

where Γ, E, M are defined in (S-36) and (S-38). Since $s = -H^{-1}\nabla$ it follows that the Hessian is equal to

$$H = M(I - E)(I - \Gamma). \tag{S-40}$$

In the case of the LRBT algorithm, if we define $\tilde{M}_t^{-1} = \tilde{D}_t^{-1} - L_t\Sigma_t L_t^T$ and $\tilde{\Gamma}_t = \tilde{M}_t^{-1}E_t^T$, we have

$$\begin{aligned} \tilde{\boldsymbol{q}}_t &= -\tilde{M}_t^{-1}(\nabla_t - E_{t-1}^T\tilde{\boldsymbol{q}}_{t-1}) \\ \tilde{\mathbf{s}}_t &= \tilde{\boldsymbol{q}}_t + \tilde{\Gamma}_t\tilde{\mathbf{s}}_{t+1}. \end{aligned} \tag{S-41}$$

Therefore, an equivalent representation holds in the sense that

$$\tilde{\boldsymbol{s}} = -\tilde{H}^{-1}\nabla, \quad \text{with} \quad \tilde{H} = \tilde{M}(I - \tilde{E})(I - \tilde{\Gamma}), \tag{S-42}$$

where the block matrices $\tilde{M}, \tilde{E}, \tilde{\Gamma}$ are defined in the same way as their exact counterparts M, E, Γ. A direct calculation shows that

$$\tilde{M}(I - \tilde{\Gamma}) = (\tilde{M}(I - \tilde{E}))^T \tag{S-43}$$

and the approximate Hessian can be written as

$$\tilde{H} = (\tilde{M}(I - \tilde{E}))\tilde{M}^{-1}(\tilde{M}(I - \tilde{E}))^T \tag{S-44}$$

which is equal to (33) in the main text. (S-44) implies that $\tilde{H}$ is positive definite (PD), if the matrices $\tilde{M}_t$ are also PD.

Lemma C.1. *The matrices $\tilde{D}_t, t = 1, \ldots, T$ are PD.*

Proof. From the recursion of $\tilde{D}_t$ ((28) in the main text), we have that when A, V_0 and V

commute and A is stable,

$$\tilde{D}_t = V^{-1}\left(A^TA + \left(\sum_{k=0}^{t-2}(A^TA)^k + V_0V^{-1}(A^TA)^{t-1}\right)^{-1}\right), \tag{S-45}$$

which is PD, by stability of A. The result holds also when the matrices do not commute, although the formulas are more complicated. □

Lemma C.2. *The matrices $\tilde{M}_t, t = 1, \ldots, T$ are PD for any choice of the threshold θ.*

Proof. We introduce the matrices $\hat{M}_t$, defined as follows:

$$\begin{aligned}\hat{M}_1 &= M_1 \\ \hat{M}_t &= D_t + B_t^TW_t^{-1}B_t - E_{t-1}\tilde{M}_{t-1}^{-1}E_{t-1}^T.\end{aligned} \tag{S-46}$$

These matrices are the matrices obtained from the exact BT recursion $M_t = D_t + B_t^TW_t^{-1}B_t - E_{t-1}M_{t-1}^{-1}E_{t-1}^T$, applied to the approximate matrices $\tilde{M}_{t-1}^{-1}$. Using (28) and (29) from the main text we can write $\hat{M}_t$ as

$$\hat{M}_t = \tilde{D}_t + B_t^TW_t^{-1}B_t + E_{t-1}L_{t-1}\Sigma_{t-1}L_{t-1}^TE_{t-1}^T = \tilde{D}_t + O_tQ_tO_t^T. \tag{S-47}$$

Using (S-47) we see that $\hat{M}_t$ is the sum of a PD matrix ($\tilde{D}_t$), and two semipositive definite (SPD) matrices (Σ_t is always PD by definition). Therefore, $\hat{M}_t^{-1}$ is also PD and equals

$$\hat{M}_t^{-1} = \tilde{D}_t^{-1} - \underbrace{\tilde{D}_t^{-1}O_t(Q_t^{-1} + O_t^T\tilde{D}_t^{-1}O_t)^{-1}O_t^T\tilde{D}_t^{-1}}_{G_t}. \tag{S-48}$$

Now $\tilde{M}_t^{-1}$ is obtained by the low rank approximation of G_t. We can write the SVD of G_t as

$$G_t = [\begin{array}{cc} L_t & R_t \end{array}]\text{blkdiag}\{\Sigma_t, S_t\}[\begin{array}{cc} L_t & R_t \end{array}]^T, \tag{S-49}$$

and have that

$$\tilde{M}_t^{-1} - \hat{M}_t^{-1} = R_tS_tR_t^T \tag{S-50}$$

Consequently $\tilde{M}_t^{-1}$ is the sum of a PD and a SPD matrix and thus is PD and so is $\tilde{M}_t$. $\square$

We now quantify the approximation error of the Hessian. It turns out that $H - \tilde{H}$ is a block diagonal matrix, which simplifies our analysis. Using (33) from the main text, we see that $\tilde{H}$ differs from H only in the diagonal blocks. From (S-46) and the Block-Thomas recursion we have that

$$\hat{M}_t + E_{t-1}\tilde{M}_{t-1}^{-1}E_{t-1}^T = D_t + B_t^T W_t^{-1} B_t = M_t + E_{t-1}M_{t-1}^{-1}E_{t-1}^T \tag{S-51}$$

and therefore by adding and subtracting $\hat{M}_t$ from the diagonal blocks of (33) we get that

$$\tilde{H} = H + \text{blkdiag}\{\tilde{M}_1 - \hat{M}_1, \ldots, \tilde{M}_T - \hat{M}_T\}. \tag{S-52}$$

This makes some intuitive sense, because the low rank approximations are applied only to the matrices that are formed from the measurements. In other words, the low rank approximation does not throw away information about the state transition dynamics. It only throws away information about the measurements, which corresponds to the entries of the main block-diagonal of the Hessian.

Using (S-52), we can easily derive bounds on the error of our approximate solution $\tilde{\mathbf{s}}$. From (S-49), we have that $\tilde{M}_t^{-1}$ is obtained by truncating the term $R_t^T S_t R_t$ of the SVD of G_t such that

$$\|L_t\Sigma_t^{1/2}\|_F^2 \geq \theta\|G_t^{1/2}\|_F^2 \Rightarrow \|R_t S_t^{1/2}\|_F^2 \leq (1-\theta)\|G_t^{1/2}\|_F^2 \Rightarrow \|R_t S_t R_t^T\| \leq (1-\theta)\|G_t^{1/2}\|_F^2, \tag{S-53}$$

and

$$\hat{M}_t^{-1} - \tilde{M}_t^{-1} = -R_t S_t R_t^T \Rightarrow \|\hat{M}_t^{-1} - \tilde{M}_t^{-1}\| \leq (1-\theta)\|G_t^{1/2}\|_F^2. \tag{S-54}$$

Using the Taylor approximation of (S-2) we have

$$\hat{M}_t = (\tilde{M}_t^{-1} - R_t S_t R_t^T)^{-1} = \tilde{M}_t + \tilde{M}_t R_t S_t R_t^T \tilde{M}_t + O((1-\theta)^2). \tag{S-55}$$

By taking the spectral norm we have

$$\|\hat{M}_t - \tilde{M}_t\| \le \|\tilde{M}_t\|^2 \|R_t S_t R_t^T\| \le (1-\theta)\|\tilde{M}_t\|^2 \|G_t^{1/2}\|_F^2, \tag{S-56}$$

We can similarly apply (S-2) to obtain bounds for $\|H^{-1} - \tilde{H}^{-1}\|$. By using (S-52) we get

$$\begin{aligned} &\|H^{-1} - \tilde{H}^{-1}\| \approx \|H^{-1}\|^2 \|\tilde{M} - \hat{M}\| \le (1-\theta) \underbrace{\|H^{-1}\|^2 \max_t \left\{ \|\tilde{M}_t\|^2 \|G_t^{1/2}\|_F^2 \right\}}_{\Psi} \Rightarrow \\ &\qquad \|\tilde{\mathbf{s}} - \mathbf{s}\| \le (1-\theta)\Psi \left\| \nabla\big|_{\mathbf{x}=0} \right\|. \end{aligned} \tag{S-57}$$

Note that $\tilde{M}_t$, G_t and thus Ψ, also depend on θ. However, this dependence is weak and practically does not affect the approximation error bounds. In fact, since $\tilde{M}_t$ and G_t are obtained from a series of low rank approximations we expect that the following relations hold

$$\begin{aligned} \|\tilde{M}_t\| &\approx \|M_t\| \\ \|G_t^{1/2}\|_F^2 &\le \|Z_t^{-1/2} U_t \tilde{D}_t^{-1}\|_F^2, \end{aligned} \tag{S-58}$$

and since the right parts of (S-58) do not depend on θ, $\|H^{-1} - \tilde{H}^{-1}\|$ and $\|\tilde{\mathbf{s}} - \mathbf{s}\|$ scale both as $O(1-\theta)$.

Remark C.3. *For the $O(1-\theta)$ bound to hold, we also need to ensure that $\|\tilde{M}_t\|$ and $\|G_t\|$ do not grow indefinitely as t increases. This holds if $\|A\| < 1$, since in this case $\tilde{D}_t$ from* (S-45) *converges to a finite matrix, and consequently M_t stays bounded as can be seen from* (S-33) *and* (S-35)*. If $\|A\| = 1$, then $\tilde{D}_t$ grows without bound and our approximation result does not hold. As a result our algorithm is not applicable in this case.*

C.2 Discussion on the iterative LRBT algorithm

Since the objective function is quadratic, the step size of the iterative algorithm can be obtained in closed form. To obtain the length of the step t_n, note that the quadratic objective function can be written as

$$f(\mathbf{s}) = \frac{1}{2}\mathbf{s}^T H\mathbf{s} + \nabla_0^T \mathbf{s} + \text{const}, \tag{S-59}$$

where ∇_0 denotes the gradient of f at the zero vector. Then to determine t_n we want to minimize $f(\mathbf{s}_n + t\mathbf{s}_{\text{dir}})$ with respect to t. Ignoring the terms that do not depend on t we have

$$t_n = \arg\min_t \left\{ \frac{1}{2}(\mathbf{s}_{\text{dir}}^T H\mathbf{s}_{\text{dir}})t^2 + (\nabla_0^T \mathbf{s}_{\text{dir}} + \mathbf{s}_n^T H\mathbf{s}_{\text{dir}})t \right\} = -\frac{\nabla_0^T \mathbf{s}_{\text{dir}} + \mathbf{s}_n^T H\mathbf{s}_{\text{dir}}}{\mathbf{s}_{\text{dir}}^T H\mathbf{s}_{\text{dir}}}. \tag{S-60}$$

The algorithm is summarized below (Alg. S-1).

Algorithm S-1 Steepest Descent using the LRBT algorithm

Initialize: $\tilde{\mathbf{s}}_0 = \mathbf{0},\ \nabla_0 \triangleq \left(\nabla\big|_{X=0}\right)$.
repeat
  $\mathbf{s}_{\text{dir}} = -\tilde{H}^{-1}\left(\nabla\big|_{X=\mathbf{s}_n}\right)$. Compute using the LRBT Algorithm (Alg. 3).
  $t_n = -(\nabla_0^T \mathbf{s}_{\text{dir}} + \mathbf{s}_n^T H\mathbf{s}_{\text{dir}})/(\mathbf{s}_{\text{dir}}^T H\mathbf{s}_{\text{dir}})$. Optimal step size.
  $\mathbf{s}_{n+1} = \mathbf{s}_n + t_n\mathbf{s}_{\text{dir}}$.
until Convergence criterion satisfied. (e.g. $\|H\mathbf{s}_{n+1} + \nabla_0\|/\|\nabla_0\| < \varepsilon$.)

By denoting $\nabla_n = \nabla\big|_{X=\mathbf{s}_n}$ and noting that $\nabla_n = H\mathbf{s}_n + \nabla_0$, t_n can be rewritten as

$$t_n = -\frac{\nabla_n^T \mathbf{s}_{\text{dir}}}{\mathbf{s}_{\text{dir}}^T H\mathbf{s}_{\text{dir}}} = \frac{\nabla_n^T \tilde{H}^{-1}\nabla_n}{\nabla_n^T \tilde{H}^{-1} H \tilde{H}^{-1}\nabla_n}. \tag{S-61}$$

Since $\tilde{H}$ approximates H, we see that the algorithm will take almost full steps of the order $1 - O(1-\theta)$. Moreover, the convergence rate of this steepest descent algorithm is linear and in general we have

$$f(\mathbf{s}_n) - f^* \leq (1 - 1/\kappa(\tilde{H}^{-1/2} H \tilde{H}^{-1/2}))(f(\mathbf{s}_{n-1}) - f^*), \tag{S-62}$$

where $\kappa(\cdot)$ denotes the condition number (Boyd and Vandenberghe, 2004). Theorem 4.1 and our simulations indicate that $1 - 1/\kappa(\tilde{H}^{-1/2} H \tilde{H}^{-1/2}) = O(1-\theta)$ and therefore

$$f(\mathbf{s}_n) - f^* \propto \gamma_\theta^n, \;\; \text{with} \;\; \gamma_\theta = O(1-\theta). \tag{S-63}$$

D Covariance estimation using the LRBT smoother

The LRBT algorithm provides only an approximate estimate of the smoothed mean $\mathbb{E}(x_t|Y_{1:T})$. However, with a few modifications, we can also use it to provide an estimate of the smoothed covariance $C_t^s = \mathrm{Cov}(x_t|Y_{1:T})$ as well, again with complexity that scales linearly with the state dimension d. To do that we can adapt to our setting the algorithm of Rybicki and Hummer (1991) for the fast solution for the diagonal elements of the inverse of a tridiagonal matrix. In the exact case, Alg. S-2 shows the modifications of the BT algorithm to give C_t^s.

Algorithm S-2 Covariance Estimation with the Block-Thomas Algorithm

$M_1 = D_1 + B_1^T W_1^{-1} B_1$ (cost $O(d^3)$)
for $t = 2$ to T **do**
  $M_t = D_t + B_t^T W_t^{-1} B_t - E_{t-1} M_{t-1}^{-1} E_{t-1}^T$ (cost $O(d^3)$)
$N_T = D_T + B_T^T W_T^{-1} B_T$ (cost $O(d^3)$)
for $t = T-1$ to 2 **do**
  $N_t = D_t + B_t^T W_t^{-1} B_t - E_t^T N_{t+1}^{-1} E_t$ (cost $O(d^3)$)
for $t = 1$ to $T-1$ **do**
  $C_t^s = (I_d - M_t^{-1} E_t N_{t+1}^{-1} E_t^T)^{-1} M_t^{-1}$ (cost $O(d^3)$)
$C_T^s = M_T^{-1}$

Compared to the exact BT algorithm, Alg. S-2 adds an additional backwards recursion that constructs the sequence of the matrices N_t defined as

$$\begin{aligned} N_T &= D_T + B_T^T W_T^{-1} B_T \\ N_t &= D_t + B_t^T W_t^{-1} B_t - E_t^T N_{t+1}^{-1} E_t, \quad t = T-1, \dots, 1. \end{aligned} \tag{S-64}$$

It is easy to see the analogy between the matrices N_t and M_t; N_t are the matrices constructed from the BT smoother when run backwards in time. Consequently, similar to (29) in the main text the matrices N_t^{-1} can be approximated as

$$N_t^{-1} \approx \tilde{N}_t^{-1} = (\tilde{D}_t^b)^{-1} - L_t^b \Sigma_t^b (L_t^b)^T, \tag{S-65}$$

where $\tilde{D}_t^b$ is matrix similar to $\tilde{D}_t$ that can be enables fast computations and (backwards) updating, and the term $L_t^b \Sigma_t^b (L_t^b)^T$ acts as a low rank perturbation. With that in mind, it is easy to modify the LRBT algorithm to also produce approximations of the smoothed covariance $C_t^s = \mathrm{Cov}(x_t | Y_{1:T})$, while still operating with $O(d)$ complexity. Pseudocode for this algorithm is given below (Alg. S-3). We leave the details as well as a theoretical analysis of this approximation method for future work, but we provide an implementation in our accompanying code.

References

Abramowitz, M. and I. Stegun (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization.* Oxford University Press.

Brown, E., L. Frank, D. Tang, M. Quirk, and M. Wilson (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience 18*, 7411–7425.

Dembo, R., S. Eisenstat, and T. Steihaug (1982). Inexact Newton methods. *SIAM Journal on Numerical analysis 19*(2), 400–408.

Eisenstat, S. and H. Walker (1994). Globally convergent inexact Newton methods. *SIAM Journal on Optimization 4*(2), 393–422.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer.

Jungbacker, B. and S. Koopman (2007). Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika 94*, 827–839.

Algorithm S-3 Covariance Estimation with the LRBT Algorithm

$\tilde{D}_1 = D_1, L_1 = D_1^{-1} B_1^T$
$\Sigma_1 = (W_1 + B_1 D_1^{-1} B_1^T)^{-1}$
for $t = 2$ to T **do**
  $\tilde{D}_t = D_t - E_{t-1} \tilde{D}_{t-1}^{-1} E_{t-1}^T$
  $O_t = [B_t^T \quad E_{t-1} L_{t-1}], \quad Q_t = \text{blkdiag}\{W_t^{-1}, \Sigma_{t-1}\}$
  $[\hat{L}_t, \hat{\Sigma}_t^{1/2}] = \text{svd}(\tilde{D}_t^{-1} O_t (Q_t^{-1} + O_t^T \tilde{D}_t^{-1} O_t)^{-1/2})$
  Truncate $\hat{L}_t$ and $\hat{\Sigma}_t$ to L_t and Σ_t. (effective rank $k_t \leq b_t + k_{t-1} \ll d$)
$\tilde{D}_T^b = D_T$, $L_T^b = (\tilde{D}_T^b)^{-1} B_T^T$
$\Sigma_1^b = (W_T + B_T D_T^{-1} B_T^T)^{-1}$
for $t = T-1$ to 2 **do**
  $\tilde{D}_t^b = D_t - E_{t+1}^T (\tilde{D}_{t+1}^b)^{-1} E_{t+1}$
  $O_t = [B_t^T \quad E_{t+1}^T L_{t+1}^b], \quad Q_t = \text{blkdiag}\{W_t^{-1}, \Sigma_{t+1}^b\}$
  $[\hat{L}_t^b, (\hat{\Sigma}_t^b)^{1/2}] = \text{svd}((\tilde{D}_t^b)^{-1} O_t (Q_t^{-1} + O_t^T (\tilde{D}_t^b)^{-1} O_t)^{-1/2})$
  Truncate $\hat{L}_t^b$ and $\hat{\Sigma}_t^b$ to L_t^b and Σ_t^b. (effective rank $k_t^b \leq b_t + k_{t+1}^b \ll d$)
for $t = 1$ to $T-1$ **do**
  $J_t = I_d - \tilde{D}_t^{-1} E_t (\tilde{D}_{t+1}^b)^{-1} E_{t+1}^T$
  $A_1 = [L_t \Sigma_t, \quad \tilde{D}_t^{-1} E_t L_{t+1}^b \Sigma_{t+1}^b - L_t \Sigma_t L_t^T E_t L_{t+1}^b \Sigma_{t+1}^b]$
  $A_2 = [E_{t+1} (\tilde{D}_{t+1}^b)^{-1} E_t^T L_t, \quad E_{t+1} L_{t+1}^b]^T$
  $A_3 = (I + A_2^T J_t^{-1} A_1)^{-1} A_2^T J_t^{-1}$
  $B_1 = J_t^{-1} [A_1, \quad L_t]$
  $B_2 = [\tilde{D}_t^{-1} A_3^T - L_t \Sigma_t L_t^T A_3^T, \quad L_t \Sigma_t]$
  $[Q_1, R_1] = \text{qr}(B_1, 0)$
  $[Q_2, R_2] = \text{qr}(B_2, 0)$
  $[U, S_t] = \text{svd}(R_1 R_2^T)$
  $P_t = Q_1 U$
  Store J_t, P_t, S_t. ($\tilde{C}_t^s = J_t^{-1} \tilde{D}_t^{-1} - P_t S_t P_t^T$)
$C_T^s = \tilde{D}_T^{-1} - L_T \Sigma_T L_T^T$.

Paninski, L., Y. Ahmadian, D. Ferreira, S. Koyama, K. Rahnama, M. Vidne, J. Vogelstein, and W. Wu (2010). A new look at state-space models for neural data. *Journal of Computational Neuroscience 29*, 107–126.

Papandreou, G. and A. Yuille (2010). Gaussian sampling by local perturbations. In *Proc. NIPS.*

Pnevmatikakis, E. and L. Paninski (2012). Fast interior-point inference in high-dimensional, sparse, penalized state-space models. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS). J Mach Learn Res*, Volume 22, pp. 895–904.

Rybicki, G. and D. Hummer (1991). An accelerated lambda iteration method for multilevel radiative transfer. I-Non-overlapping lines with background continuum. *Astronomy and Astrophysics 245*, 171–181.

Sun, W. and Y. Yuan (2006). *Optimization Theory and Methods: Nonlinear Programming.* Springer Verlag.

Wolfram Research, Inc. (2010). *Mathematica Edition: Version 8.0.* Champaign, IL: Wolfram Research, Inc.