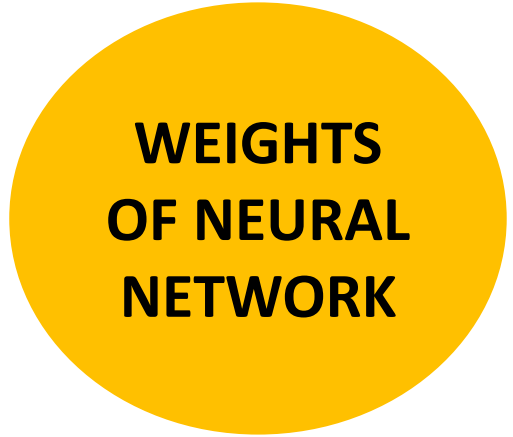
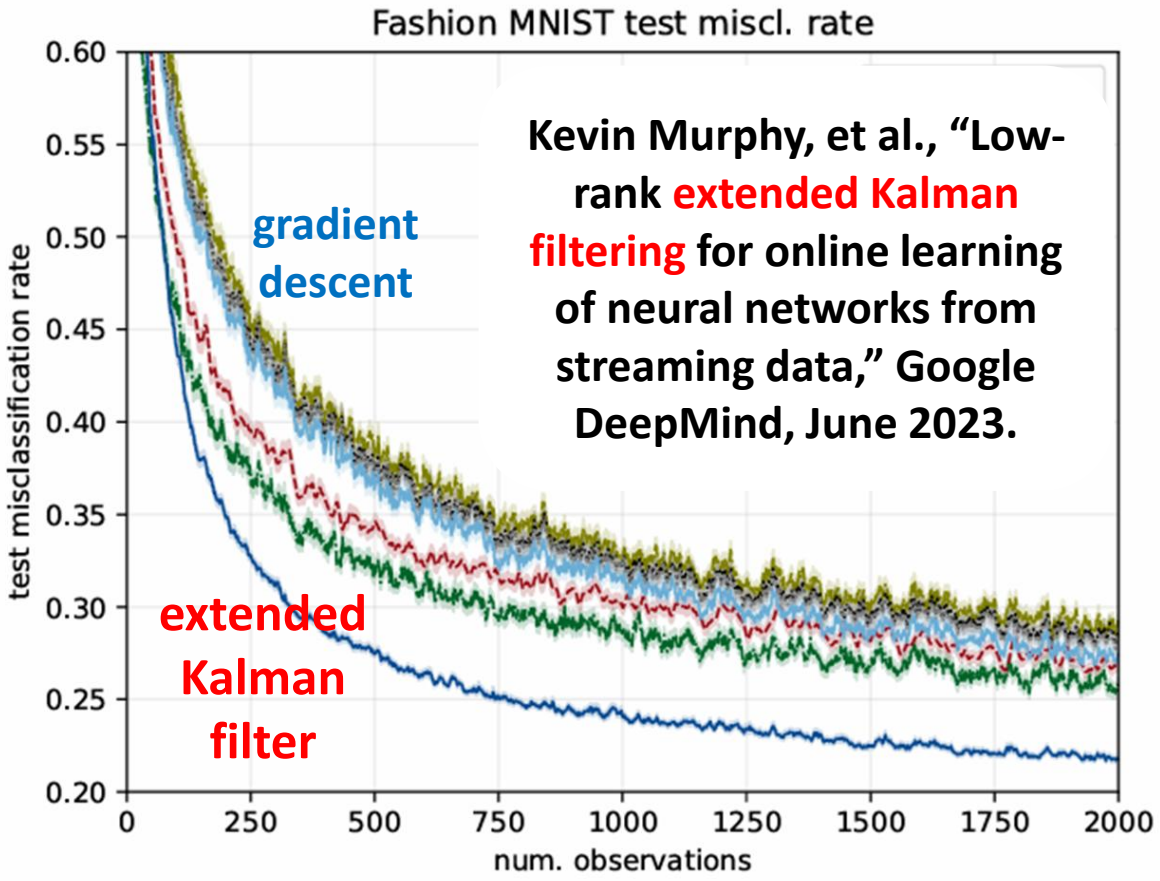
Raytheon

fast optimal Kalman filter

Fred Daum

14 June 2026



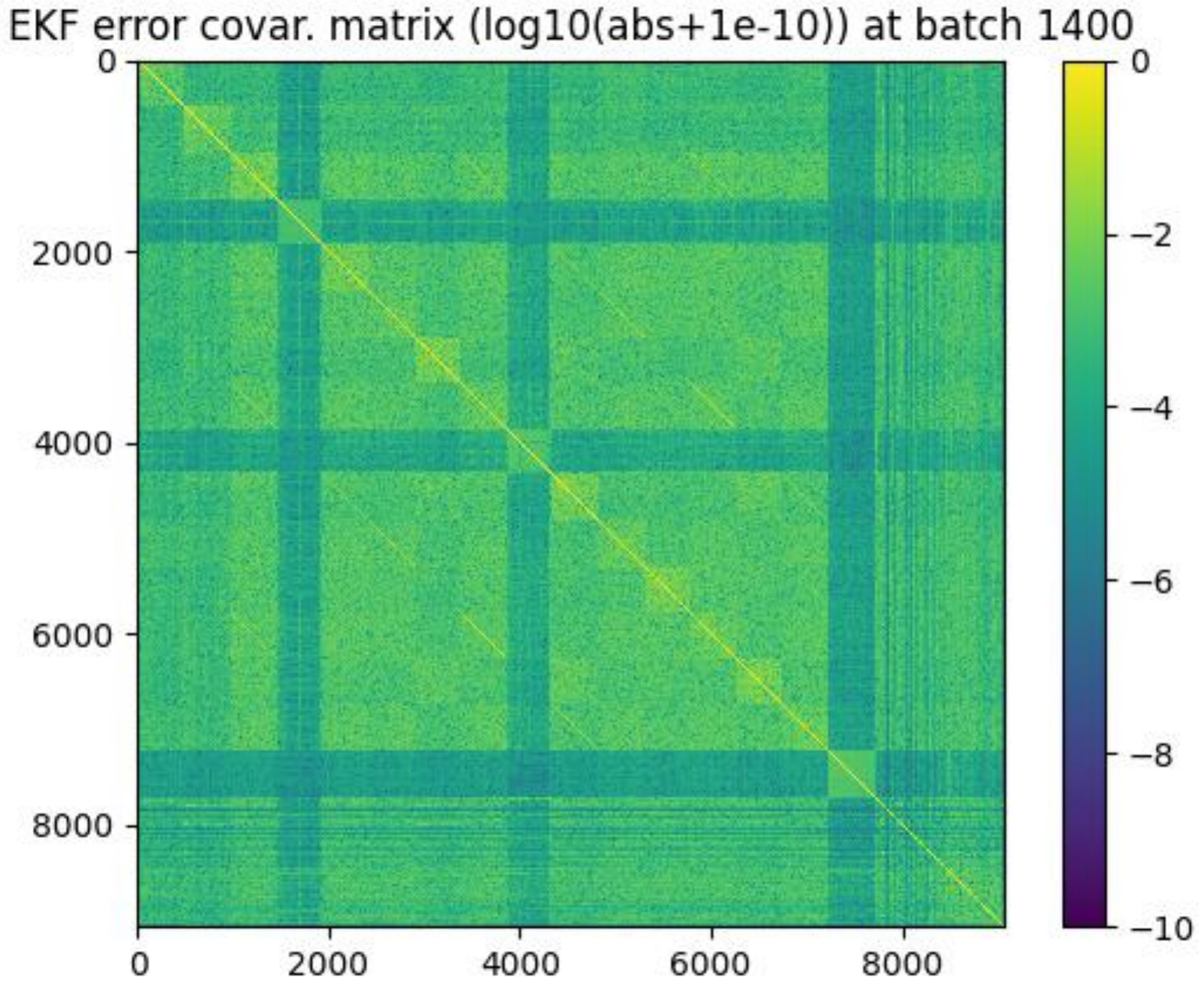


We want to invent a new Kalman filter like algorithm that gives optimal estimation accuracy subject to the constraint that the memory and computer run time scales linearly with the dimension of the state vector.

number of parameters in model	quadratic scaling cost of SRAM	linear scaling cost of SRAM	quadratic scaling cost of DRAM	linear scaling cost of DRAM
trillion (SOTA LLMs today)	$28 million trillions	$28 million	$64 thousand trillions	$64 thousand
billion	$28 trillion	$28 thousand	$64 billion	$64
million	$28 million	$28	$64 thousand	6 cents
thousand	$28	3 cents	6 cents	1 cent

quadratic scaling of GPU memory with model size is the kiss of death for deep learning

covariance matrix approximation	comments	references
1. diagonal	**highly suboptimal accuracy for Kalman filters** (Murphy & our numerical experiments)	Maybeck, Feldkamp, Murphy, Fernandez
2. low-rank product	fast, but **bad for Kalman filters (unstable)**	AI papers for Hessian
3. block diagonal	**does not work well for transformer models**	see our plots
4. **diagonal plus low-rank product D + VV***	**computational complexity that is linear in d, with much better accuracy than diagonal**	Murphy (2023) & Bonnabel (2024) et al.
5. k-diagonal	not good for deep learning	numerical PDE papers
6. Kronecker products, Hankel-Toeplitz, etc.	extensive literature of structured matrices for many diverse applications (e.g., deep learning)	Martens, Pan, Kailath, Hero, et al.
7. random projections	for sample covariance matrices (not Kalman)	Bengtsson, Bickel, Cai
8. random zeros in covariance matrix	vaguely like Hinton's dropout; no theory, but great in many real world applications	Musco for PSD (2019), Go & Pan (2025)
9. **OPTIMAL APPROXIMATIONS**	hopefully invented this week at SIAM MPI!	none known



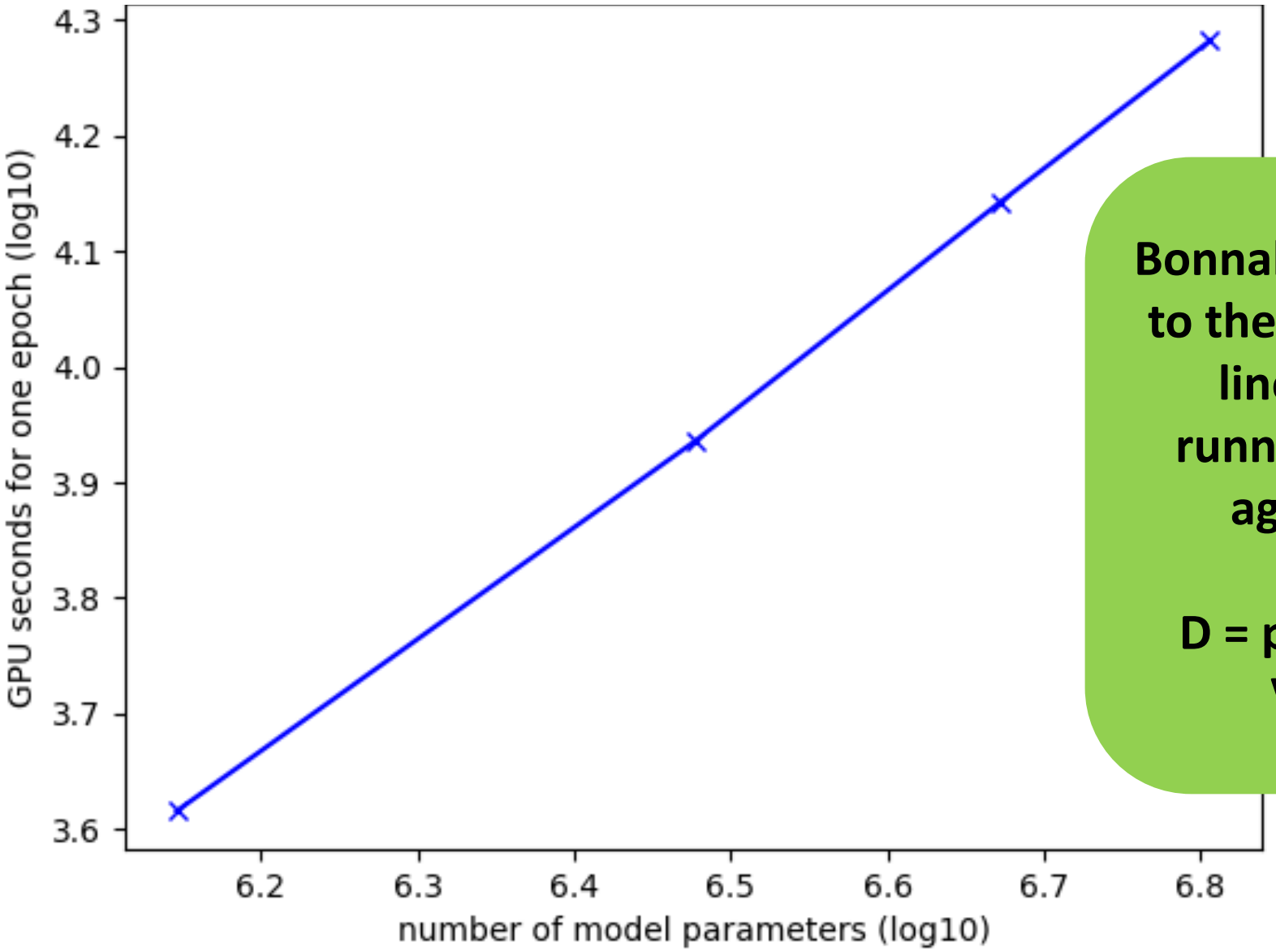
the state vector error covariance matrix for the Kalman filter is not well approximated by a diagonal matrix for deep learning

Low-rank plus diagonal approximations for Riccati-like matrix differential equations

Silvère Bonnabel
MINES ParisTech, PSL University, Center for robotics
silvere.bonnabel@mines-paristech.fr

Abstract

We consider the problem of computing tractable approximations of time-dependent $d \times d$ large positive semi-definite (PSD) matrices defined as solutions of a matrix differential equation. We propose to use "low-rank plus diagonal" PSD matrices as approximations that can be stored with a memory cost being linear in the high dimension d. To constrain the solution of the differential equation to remain in that subset, we project the derivative at all times onto the tangent space to the subset, following the methodology of dynamical low-rank approximation. We derive a closed-form formula for the projection, and show that after some manipulations it can be computed with a numerical cost being linear in d, allowing for tractable implementation. Contrary to previous approaches based on pure low-rank approximations, the addition of the diagonal term allows for our approximations to be invertible matrices, that can moreover be inverted with linear cost in d. We apply the technique to Riccati-like equations, then to two particular problems. Firstly a low-rank approximation to our recent Wasserstein gradient flow for Gaussian approximation of posterior distributions in approximate Bayesian inference, and secondly a novel low-rank approximation of the Kalman filter for high-dimensional systems. Numerical simulations illustrate the results.

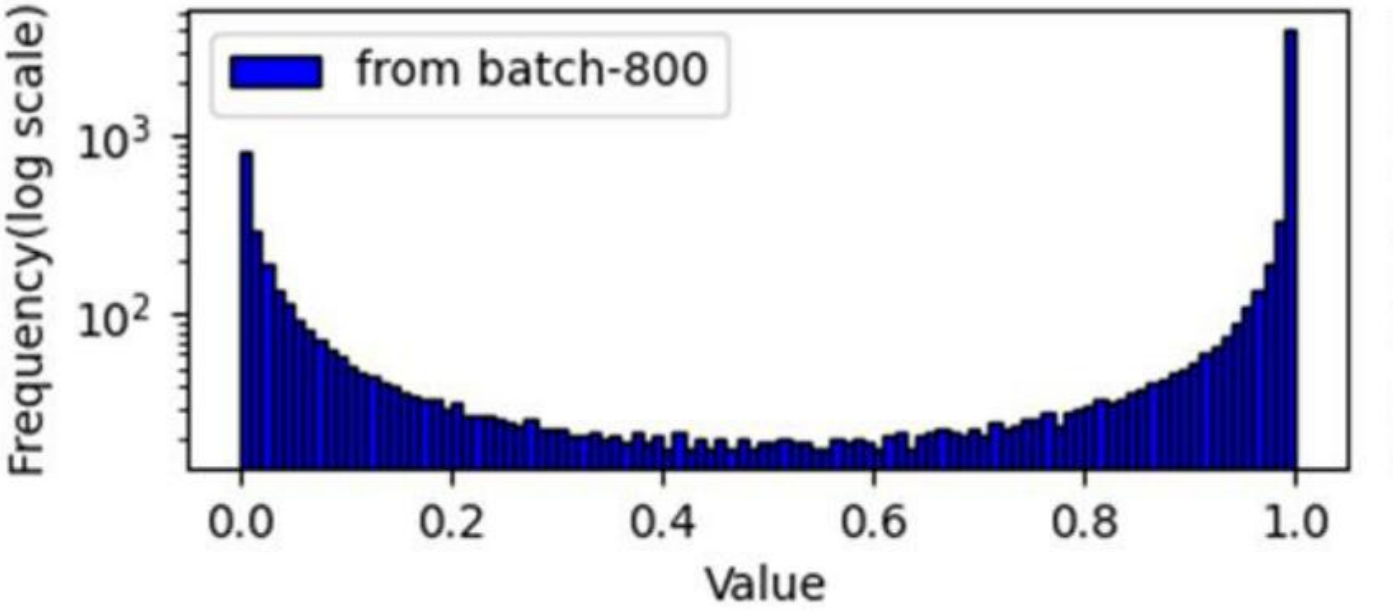


Bonnabel's PPCA approximation to the covariance matrix scales linearly with model size, running on a GPU, in perfect agreement with theory

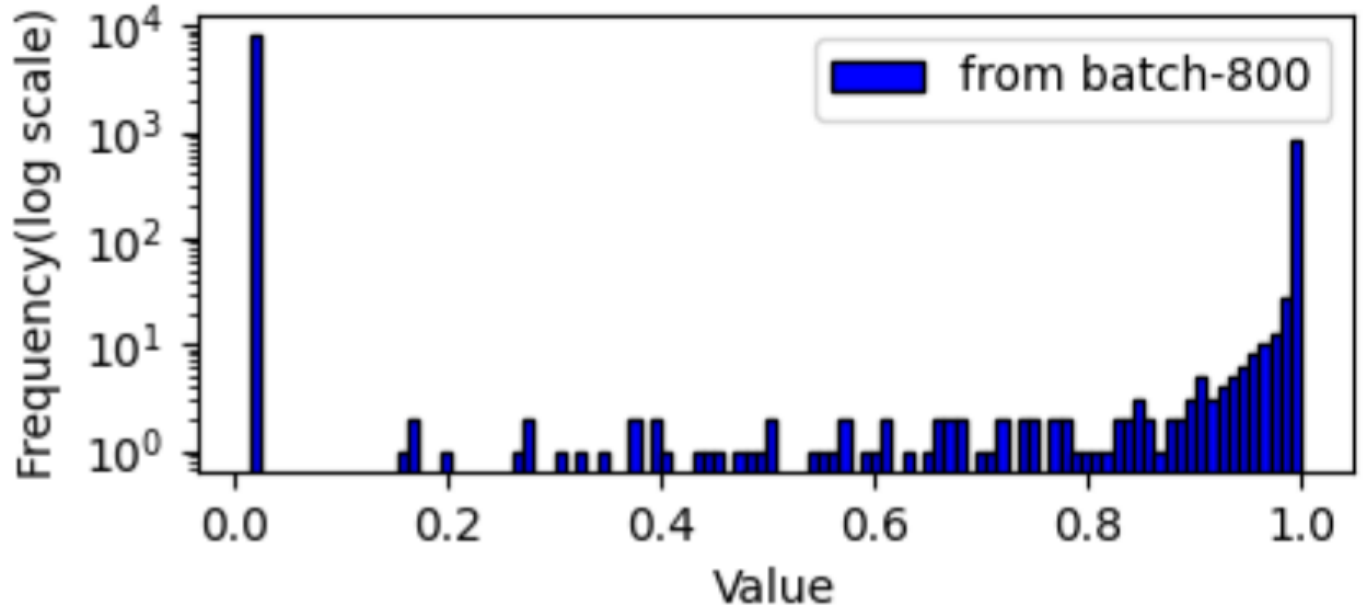
$P \approx D + VV'$

D = positive diagonal matrix

V = low rank matrix



spectrum of eigenvalues of the covariance matrix for an exact Kalman filter



spectrum of eigenvalues for Kalman filter using Bonnabel's PPCA approximation to the covariance matrix
$P \approx D + VV'$

assuming that the filter has the form: $\hat{x}$ (j) = $\bar{x}$(j) + K(j) $[z(j) - H(j)\bar{x}(j)]$
we can compute the error covariance matrix of x for any K as follows:

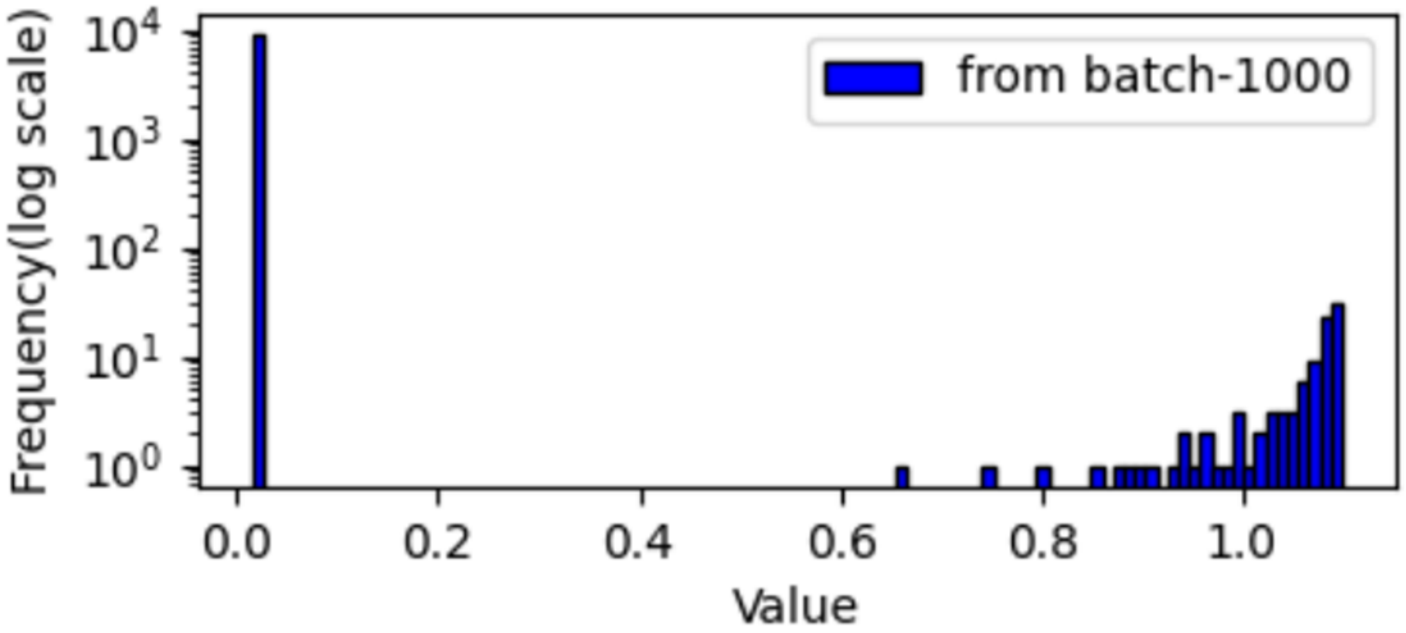
$$P = (I - KH)M(I - KH)^T + KRK^T$$ "Joseph form" of filter

the "optimal" Kalman filter gain matrix (K*) can be derived by minimizing Tr(P) by computing dTr(P)/dK = 0 or by completing the square of Tr(P) and solving for the matrix K:

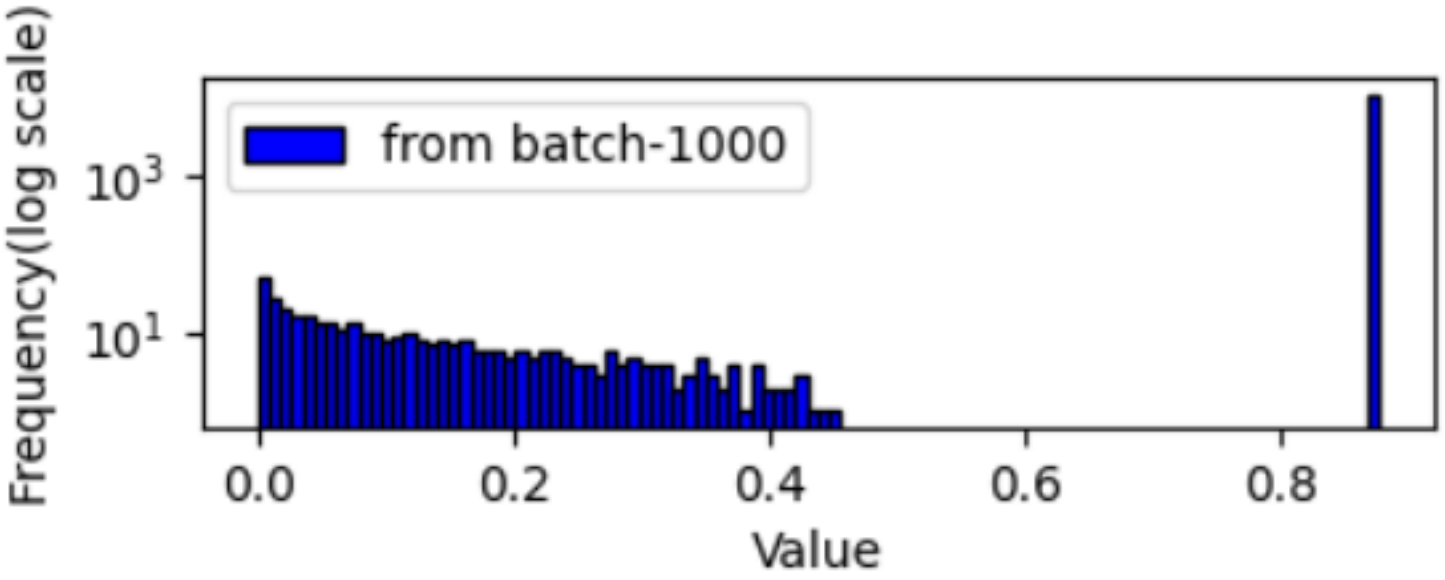
$$K^* = MH^T(HMH^T + R)^{-1}$$

K* is only really optimal if the covariance matrix M is actually correct (rather than a low-rank approximation of the true covariance)

R = measurement error covariance matrix
M = predicted state vector error covariance matrix
H = measurement sensitivity matrix, as in z = Hx + v
x = state vector to be estimated
v = Gaussian zero mean measurement error with covariance matrix R



Bonnabel PPCA without Joseph form of EKF



Bonnabel PPCA with Joseph form of EKF has much more accurate spectrum for smallest eigenvalues

LOW-RANK EXTENDED KALMAN FILTERING FOR ONLINE LEARNING OF NEURAL NETWORKS FROM STREAMING DATA

Peter G. Chang
U. Chicago

Gerardo Durán-Martín
Queen Mary Univ.

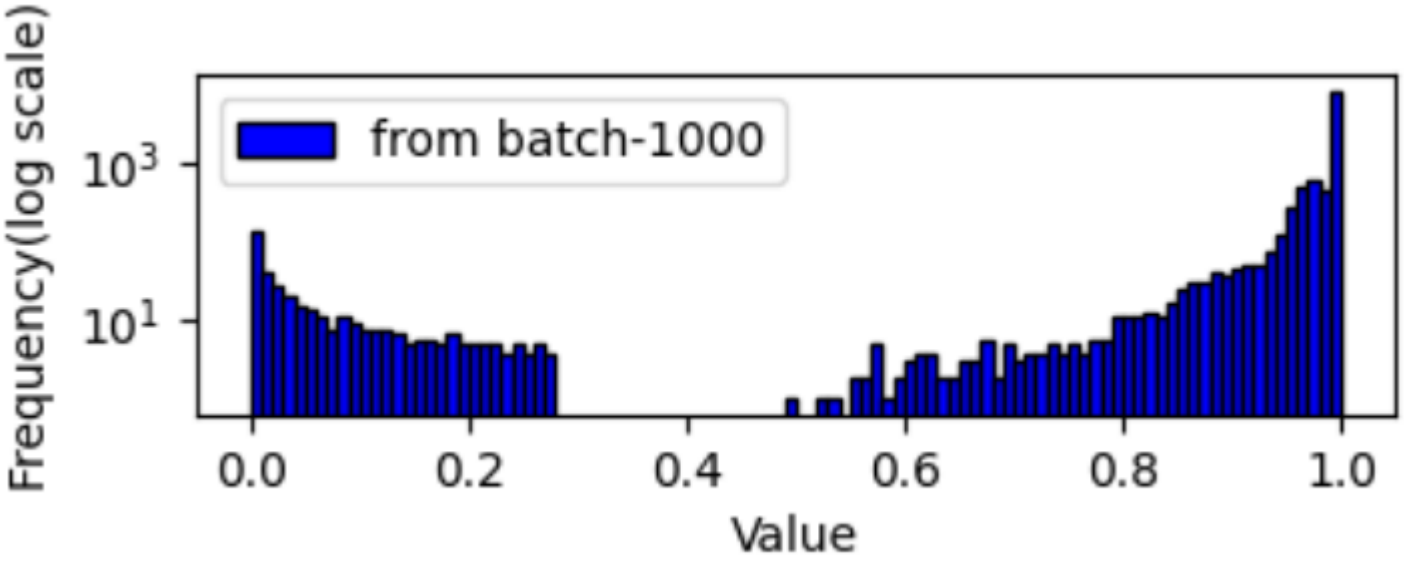
Alex Shestopaloff
Queen Mary Univ.

Matt Jones
U. Colorado, Boulder

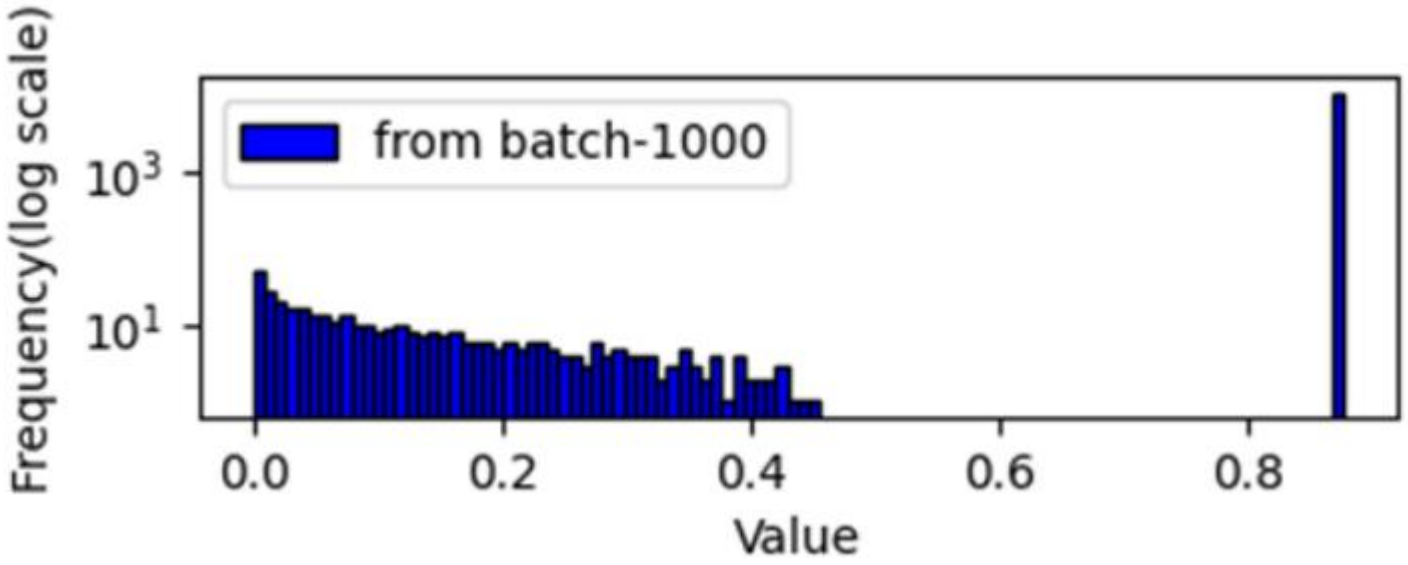
Kevin Murphy
Google DeepMind

ABSTRACT

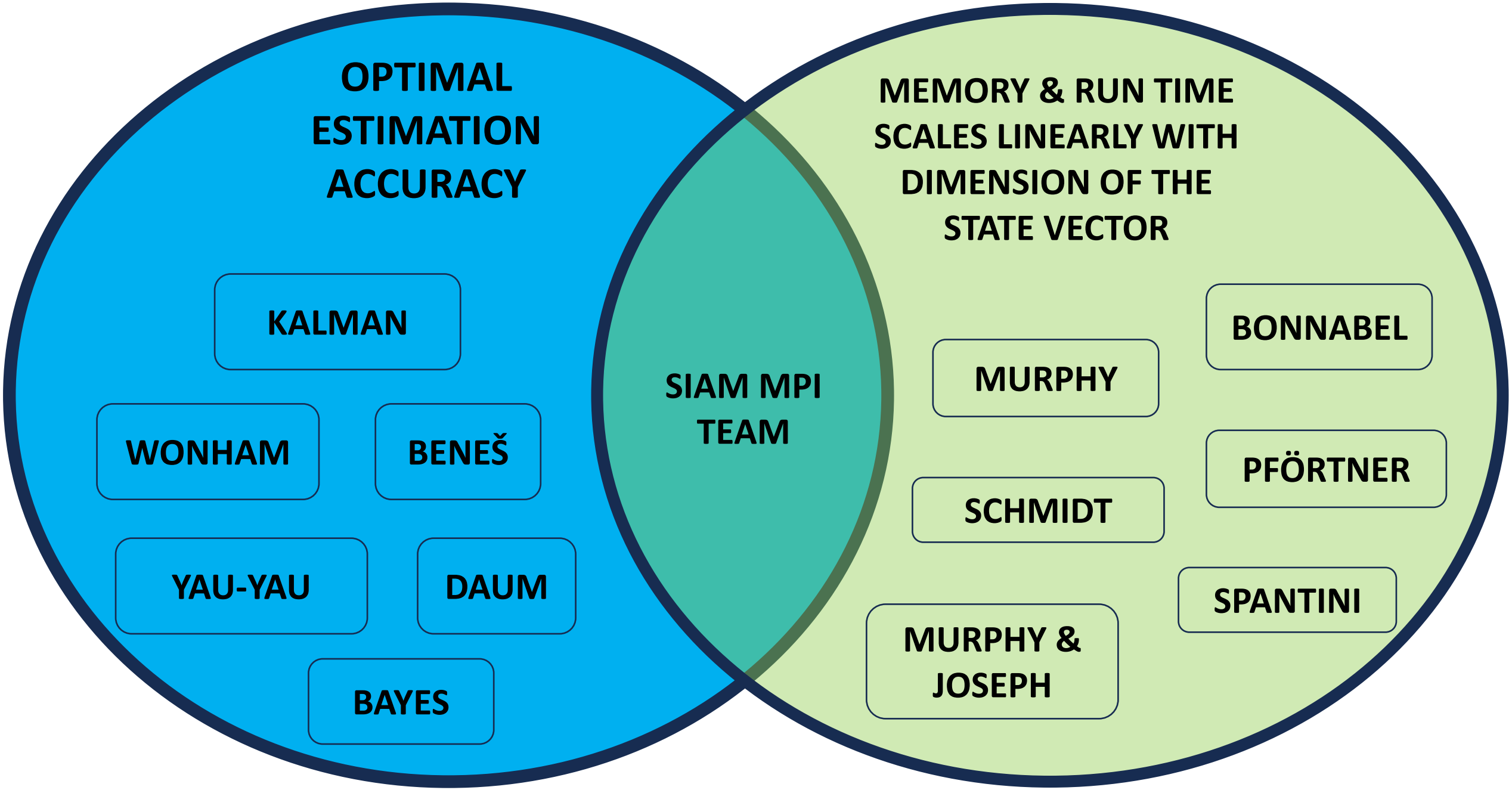
We propose an efficient online approximate Bayesian inference algorithm for estimating the parameters of a nonlinear function from a potentially non-stationary data stream. The method is based on the extended Kalman filter (EKF), but uses a novel low-rank plus diagonal decomposition of the posterior precision matrix, which gives a cost per step which is linear in the number of model parameters. In contrast to methods based on stochastic variational inference, our method is fully deterministic, and does not require step-size tuning. We show experimentally that this results in much faster (more sample efficient) learning, which results in more rapid adaptation to changing distributions, and faster accumulation of reward when used as part of a contextual bandit algorithm.



Murphy's method with Joseph form of EKF
P^{-1}≈ D +VV'



Bonnabel PPCA with Joseph form of EKF
P ≈ D +VV'


OPTIMAL ESTIMATION ACCURACY
KALMAN
WONHAM
BENEŠ
YAU-YAU
DAUM
BAYES
SIAM MPI TEAM
MEMORY & RUN TIME SCALES LINEARLY WITH DIMENSION OF THE STATE VECTOR
BONNABEL
MURPHY
PFÖRTNER
SCHMIDT
SPANTINI
MURPHY & JOSEPH

BACKUP

WARNING:

To approximate covariance matrices, it is tempting to minimize the Frobenius norm of the error (and more general norms, e.g., variational Schatten-p quasi-norms), but this is a bad idea for Kalman filters for four reasons:

(1) it does not guarantee that the approximate matrix is positive definite, and hence the Kalman filter might be unstable.

(2) it is computationally extremely expensive.

(3) it does not exploit the structure of the Kalman filter problem (e.g., H & K = low-rank matrices; see Spantini's paper).

(4) it values approximation accuracy for the biggest eigenvalues, which correspond to the state variables with the worst estimation accuracy, at the cost of approximation accuracy for the smallest eigenvalues, which are the most important for Kalman filters.

GIVEN:

H(j), R & P(0)

for a set of measurements:

z(j) = H(j)x(j) + v(j)

with trivial dynamics: x(j+1) = x(j)

(j = 1, 2,…,m)

Gaussian & linear problem (like the Kalman filter problem);

assume that the filter has the same form as a Kalman filter:

$$\hat{x} = \bar{x} + K\left[z - H\bar{x}\right]$$

K = Kalman filter gain matrix;

H = very low rank matrix

R & P(0) = diagonal positive definite matrices

WANT:

design a positive definite approximation to the error covariance matrix of the state vector and derive the optimal filter gain matrix K such that we get optimal estimation accuracy of the state vector subject to the constraints that memory and computer run time scale linearly with the dimension of the state vector

REFERENCES

[1] Silvère Bonnabel, Marc Lambert and Francis Bach, "Low-rank plus diagonal approximations for Riccati-like matrix differential equations," SIAM Journal 2024, https://arxiv.org/abs/2407.03373.

[2] Peter Chang, et al., "Low-Rank Extended Kalman Filtering for Online Learning of Neural Networks from Streaming Data," 2023, https://arxiv.org/abs/2305.19535.

[3] Jonathan Schmidt, et al., "The Rank-Reduced Kalman Filter: Approximate Dynamical-Low-Rank Filtering In High Dimensions," NIPS 2023, https://arxiv.org/abs/2306.07774.

[4] James Martens, et al., "Optimizing Neural Networks with Kronecker-factored Approximate Curvature," June 2020, https://arxiv.org/abs/1503.05671.

[5] Alessio Spantini, et al., "Optimal Low-Rank Approximations of Bayesian Linear Inverse Problems," July 2015, https://arxiv.org/abs/1407.346.

[6] Marvin Pförtner, et al., "Computation-Aware Kalman Filtering and Smoothing," March 2025, https://arxiv.org/pdf/2405.08971.

	matrix approximated	approximation	computer run time d = model size r = rank of V matrix	memory
1. Kalman (1960)	covariance matrix	exact	d^3	$d^2/2$
2. Bonnabel (SIAM 2024)	covariance matrix (FA version)	diagonal plus low-rank: D + VV' (rank V = r)	$dr^2 + r^6$	$dr + r^4/4$
3. Bonnabel (SIAM 2024)	covariance matrix (PPCA version)	D + VV'	dr^2	dr
4. Murphy (2023)	information matrix	D + VV'	dr^2	$d + dr + r^2$
5. Schmidt (NIPS 2023)	square root covariance matrix	D + VV'	$dr^2 + mr^2 + r^3$	TBD
6. TBD	square root information matrix	D + VV'	dr^2	$d + dr + r^2$
7. Scheffé (2025)	Schur complement*	D + VV'	TBD	TBD
8. Kim & Daum (2023)	non-square square root (i.e., not Cholesky)	D + VV'	TBD	TBD
9. Spantini & Marzouk (2015)	information matrix	D + VV'	various	various

THEORY	CONNECTION TO OUR PROBLEM	REFERENCES
1. fixed finite dimensional **exact** optimal nonlinear filters (estimation Lie algebra for Zakai equation)	applies to Kalman filter; we want the "integrated" dimension of the filter and we want an **approximation** (not exact)	Brockett, Marcus, Mitter, Ocone, Clark, Hazewinkel, Chaleyat-Maurel
2. fixed finite dimensional **exact** optimal nonlinear filters (exponential family for Zakai & Bayes)	same as above	Daum, Yau & Yau
3. minimum cost **approximate** representation of real numbers	bits are minimum cost for given approximation accuracy (dual for us)	Wiener, von Neumann, Aiken, IBM, Intel, NVIDIA
4. information based complexity	very general theory of optimal algorithms for function **approximation** with cost	Wozniakowski, Traub, Hickernell, Sloan, Novak
5. computational complexity over reals	similar to IBC above	Smale, Blum & Shub
6. information theory	optimal encoding & decoding of information with entropy criterion	Shannon, et al.
7. compressive sensing	random projections approximation	Candes, Tao 2006
8. sufficient statistics	exponential family	Fisher, Darmois, Pitman
9. batch least squares	sufficient statistic grows linearly with data (dual of our problem)	Gauss
10. optimal stochastic control theory	cost or constraint on control vs. memory	Bellman, Kalman, Cox

Suppose we are given a vector f in a class $\mathcal{F} \subset \mathbb{R}^N$, e.g. a class of digital signals or digital images. How many linear measurements do we need to make about f to be able to recover f to within precision ϵ in the Euclidean (ℓ_2) metric?

This paper shows that if the objects of interest are sparse in a fixed basis or compressible, then it is possible to reconstruct f to within very high accuracy from a small number of random measurements by solving a simple linear program. More precisely, suppose that the nth largest entry of the vector $|f|$ (or of its coefficients in a fixed basis) obeys $|f|_{(n)} \leq R \cdot n^{-1/p}$, where $R > 0$ and $p > 0$. Suppose that we take measurements $y_k = \langle f, X_k \rangle$, $k = 1, \ldots, K$, where the X_k are N-dimensional Gaussian vectors with independent standard normal entries. Then for each f obeying the decay estimate above for some $0 < p < 1$ and with overwhelming probability, our reconstruction $f^\sharp$, defined as the solution to the constraints $y_k = \langle f^\sharp, X_k \rangle$ with minimal ℓ_1 norm, obeys

$$\|f - f^\sharp\|_{\ell_2} \leq C_p \cdot R \cdot (K/\log N)^{-r}, \quad r = 1/p - 1/2.$$

There is a sense in which this result is optimal; it is generally impossible to obtain a higher accuracy from any set of K measurements whatsoever. The methodology