Numerical Algorithms
https://doi.org/10.1007/s11075-022-01332-9

ORIGINAL PAPER

Iterate averaging, the Kalman filter, and 3DVAR for linear inverse problems

Felix G. Jones[1] · **Gideon Simpson**[1]

Received: 5 December 2021 / Accepted: 4 May 2022

Abstract
It has been proposed that classical filtering methods, like the Kalman filter and 3DVAR, can be used to solve linear statistical inverse problems. In the work of Iglesias, Lin, Lu, and Stuart (*Commun. Math. Sci.* 15(7):1867–1896, 2017), error estimates were obtained for this approach. By optimally tuning a regularization parameter in the filters, the authors were able to show that the mean squared error could be systematically reduced. Building on the aforementioned work of Iglesias, Lin, Lu, and Stuart, we prove that by (i) considering the problem in a weaker norm and (ii) applying simple iterate averaging of the filter output, 3DVAR will converge in mean square, unconditionally on the choice of parameter. Without iterate averaging, 3DVAR cannot converge by running additional iterations with a fixed choice of parameter. We also establish that the Kalman filter's performance in this setting cannot be improved through iterate averaging. We illustrate our results with numerical experiments that suggest our convergence rates are sharp.

Keywords Kalman filter · 3DVAR · Statistical inverse problems · Averaging

Mathematics Subject Classification (2010) 93E11 · 65J22 · 47A52

1 Introduction

The focus of this work is on the inverse problem

$$y = Au^\dagger + \eta, \tag{1.1}$$

where, given the noisy observation y of $Au^\dagger$, we wish to infer $u^\dagger$. In our setting, $A : X \to Y$ is a compact operator between separable Hilbert spaces and $\eta \sim N(0, \gamma^2 I)$

✉ Gideon Simpson
grs53@drexel.edu

[1] Department of Mathematics, Drexel University, Philadelphia, PA, 19103, USA

is white noise, modelling measurement error. This problem is well-known to be ill-posed in the infinite-dimensional setting, as A has an unbounded inverse. Methods of solution include the use of regularized Moore-Penrose inverses and, subject to the introduction of a prior, Bayesian formulations, [4–6, 12, 19, 24, 25].

In [10], a key inspiration for the present work, Iglesias, Lin, Lu, and Stuart considered two classical filtering algorithms, the Kalman filter and 3DVAR, with the goal of using them to solve (1.1). The filtering methodology for (1.1) requires the introduction, conceptually, of the artificial dynamical system

$$u_n = u_{n-1}, \quad u_0 = u^\dagger, \tag{1.2a}$$

$$y_n = Au_n + \eta_n, \quad \eta_n \overset{\text{i.i.d.}}{\sim} N(0, \gamma^2 I). \tag{1.2b}$$

Here, at algorithmic time step n, u_n is the quantity of interest, and y_n is the noisy observation. Having ascribed a notion of time to the problem, we can then apply a filter. This provides a mechanism for estimating $u^\dagger$ in (1.1) in an online setting, where a sequence of i.i.d. observations, $\{y_n\}$, is available. This corresponds to "Data Model 1" of [10].

Amongst the key results of [10], reviewed in detail below, is that under sufficiently strong assumptions, the Kalman filter will recover the truth in mean square, unconditionally on the choice of the scalar regularization parameter. Under somewhat weaker assumptions, the error will only be bounded, though through minimax selection of a scalar parameter, an optimal error can be achieved for a given number of iterations, allowing the error to be driven to zero.

3DVAR is a simplification of Kalman that is demonstrated to have, at best, bounded error, though, again, through minimax parameter tuning, it can perform comparably to Kalman. Kalman is more expensive than 3DVAR, as it requires updating an entire covariance operator at each iteration. For finite-dimensional approximations, this may require costly matrix-matrix multiplications at each iterate.

Here, by working in a weaker norm and averaging the iterates, we are able to establish that 3DVAR will unconditionally converge in mean square for all admissible filter parameters. Such weaker convergence was also considered in [3], for a related problem on 4DVAR. Further, we show that this simple iterate averaging *cannot* improve the performance of the Kalman filter.

1.1 Filtering algorithms

The Kalman filter is a probabilistic filter that estimates a Gaussian distribution, $N(m_n, C_n)$, for $u^\dagger$ at each iterate. Given a starting mean and covariance, m_0 and C_0, the updates are as follows:

$$m_n = K_n y_n + (I - K_n A)m_{n-1}, \tag{1.3a}$$

$$C_n = (I - K_n A)C_{n-1}, \tag{1.3b}$$

$$K_n = C_{n-1}A^*(AC_{n-1}A^* + \gamma^2 I)^{-1}. \tag{1.3c}$$

Here, K_n is the so-called "Kalman gain." m_n is a point estimate of $u^\dagger$.

While Kalman is a probabilistic filter, 3DVAR is not. It is obtained by applying Kalman with a static covariance operator $C_n = \frac{\gamma^2}{\alpha}\Sigma$ for some predetermined operator Σ:

$$u_n = \mathcal{K} y_n + (I - \mathcal{K}A)u_{n-1}, \tag{1.4a}$$
$$\mathcal{K} = (A^*A + \alpha\Sigma^{-1})^{-1}A^*. \tag{1.4b}$$

We refer the reader to [7, 14, 23, 25], and references therein, for a thorough discussion and analysis of these classical filtering methods and their extensions.

Indeed, several important extensions of these classical methods that have appeared in the literature have also been directly applied to statistical inverse problems like (1.1), along with its nonlinear variation, $y = \mathcal{G}(u^\dagger) + \eta$. In particular, the ensemble Kalman filter (EnKF), using an ensemble of replicas of the problem, has been successfully applied to solve such problems in [8, 9]. See, for instance, [14, 21, 22], for additional details and analysis of EnKF. We also mention [3], which uses similar ideas with 4DVAR.

Continuous in time analogs of these methods and problems also exist, resulting in the Kalman-Bucy filter and continuous in time 3DVAR, [14, 18, 25]. In [15], these were used to solve the continuous in time analog of (1.2)

$$du = 0, \tag{1.5a}$$
$$dy = Audt + d\eta, \tag{1.5b}$$

where $\eta(t)$ is now a Weiner process in the appropriate function space, [2].

1.2 Key assumptions and prior results

In [10], the following assumptions were invoked.

Assumption 1 (1) $C_0 = \frac{\gamma^2}{\alpha}\Sigma$ with $\mathrm{Ran}(\Sigma^{\frac{1}{2}}) \subset \mathrm{Dom}(A)$, $\alpha > 0$, and Σ a self-adjoint positive definite trace class operator with Σ^{-1} densely defined.

(2) Σ induces a Hilbert scale, and there exist constants $C > 1$, $\nu > 0$ such that A induces an equivalent norm:

$$C^{-1}\|x\|_\nu \le \|Ax\| \le C\|x\|_\nu, \quad \|\bullet\|_\nu = \|\Sigma^{\frac{\nu}{2}}\bullet\|. \tag{1.6}$$

(3) The initial error is sufficiently "smooth,"

$$m_0 - u^\dagger \in \mathrm{Dom}(\Sigma^{-\frac{s}{2}}), \quad 0 \le s \le \nu + 2, \tag{1.7}$$

where we replace m_0 with u_0 in the case of 3DVAR in the above expression.

Under this first set of assumptions, Iglesias et al. established

Theorem 1.1 (Theorem 4.1 of [10]) *The Kalman filter admits the mean square error bound*

$$\mathbb{E}[\|m_n - u^\dagger\|^2] \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{s}{\nu+1}} + \frac{\gamma^2}{\alpha}\mathrm{Tr}\Sigma$$

and

Theorem 1.2 (Theorem 5.1 of [10]) *3DVAR admits the mean square error bound*

$$\mathbb{E}[\|u_n - u^\dagger\|^2] \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{s}{\nu+1}} + \frac{\gamma^2}{\alpha}\mathrm{Tr}\Sigma \log n.$$

At fixed values of α, Theorems 1.1 and 1.2 preclude convergence, and, in the case of 3DVAR, the error may even grow. However, there are two free parameters: the number of iterations n and the regularization parameter α. Indeed, within a Bayesian framework, α can be interpreted as the strength of a prior relative to a likelihood. For a fixed number of iterations, n, α can be tuned to minimize the error. Indeed, the error can be made arbitrarily small by selecting a sufficiently large n with the optimal α.

However, in both Theorems 1.1 and 1.2, there is an unknown constant. If the error at the given, optimal choice of α for a given n is inadequate, one must obtain additional data, update the value of α, and rerun the algorithm. A benefit of the present work is that, by using iterate averaging, the error of 3DVAR can always be reduced with additional iterates, without necessarily altering α and discarding previously computed iterations. We will revisit the minimax estimates under a simultaneous diagonalization assumption.

Indeed, stronger results were obtained in [10] subject to the simultaneous diagonalization assumption:

Assumption 2 (1) Σ and A^*A simultaneously diagonalize against the set $\{\varphi_i\}$ with respective eigenvalues σ_i and a_i^2, and these eigenvalues satisfy

$$\sigma_i = i^{-1-2\epsilon}, \quad a_i \asymp i^{-p}, \quad \epsilon > 0, \quad p > 0. \tag{1.8}$$

(2) $m_0 = 0$ (or u_0 in 3DVAR) and $u^\dagger$ satisfies, for $0 < \beta \leq 1 + 2\epsilon + 2p$,

$$\sum_{i=1}^{\infty} i^{2\beta} |u_i^\dagger|^2 < \infty. \tag{1.9}$$

With this, Iglesias et al. obtain

Theorem 1.3 (Theorem 4.2 of [10]) *Under Assumption 2, for the Kalman filter,*

$$\mathbb{E}[\|m_n - u^\dagger\|^2] \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{2\beta}{1+2\epsilon+2p}} + \gamma^2 n^{-\frac{2\epsilon}{1+2\epsilon+2p}} \alpha^{-\frac{1+2p}{1+2\epsilon+2p}}$$

and

Theorem 1.4 (Theorem 5.2 of [10]) *Under Assumption 2, for 3DVAR,*

$$\mathbb{E}[\|u_n - u^\dagger\|^2] \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{2\beta}{1+2\epsilon+2p}} + C\gamma^2\alpha^{-\frac{1+2p}{1+2\epsilon+2p}}.$$

Now the Kalman filter will converge at any choice of parameter, while 3DVAR has at worst a bounded error. Again, α can be tuned so as to obtain a minimax convergence rate. Indeed, in the setting where one has a fixed number of n samples, at the optimal value of α, Theorems 1.3 and 1.4 lead to the estimates (also found in [10]):

$$\mathbb{E}[\|m_n - u^\dagger\|^2] \lesssim \quad n^{-\frac{2\beta}{1+2\beta+2p}}, \tag{1.10}$$

$$\mathbb{E}[\|u_n - u^\dagger\|^2] \lesssim n^{-\frac{2\beta}{1+2\beta+2p+2\epsilon}} \log n, \tag{1.11}$$

where the first expression is for Kalman and the second is for 3DVAR. Similar expressions are also available in the general case for Theorems 1.1 and 1.2.

Thus far, we have discussed the study of problem (1.1) in a sequential setting, where the data, $\{y_n\}$, is assimilated one sample at a time. In some settings, a static, fixed, number of samples, n, may be available together. Instead of (1.1), we might then examine

$$\begin{aligned} \bar{y}_n &= Au^\dagger + \bar{\eta}_n, \qquad \bar{\eta}_n \sim N(0, \tfrac{\gamma^2}{n} I), \\ \bar{y}_n &= \tfrac{1}{n}\sum_{k=1}^{n} y_k, \qquad \bar{\eta}_n = \tfrac{1}{n}\sum_{k=1}^{n} \eta_k. \end{aligned} \tag{1.12}$$

The variance of the noise has been reduced by a factor of n. This can be solved using a regularized approximation of A^+ to obtain $\bar{u}_{n,\alpha}$. Under suitable assumptions and identifying the optimal $\alpha = \alpha_\star(n)$, one can obtain (see, for instance, [1, 12, 16, 17, 19, 26])

$$\mathbb{E}[\|\bar{u}_{n,\alpha_\star(n)} - u^\dagger\|^2] \lesssim n^{-\frac{2\beta}{1+2p+2\beta}} \tag{1.13}$$

This precisely corresponds to the minimax solution of Kalman (1.10), while there is a loss for 3DVAR (1.11). Note that this is only for the $t = 0$ norm. A generalization to the $t < 0$ norm is covered in [16] and for $t(1 + 2\epsilon) \leq 2p$ in [17]. As we are principally interested in the general $t \geq 0$ case, we state and prove our own version of theorem below using a spectral cutoff regularization.

1.3 Main results

The main results of this paper are contained in the following theorems.

First, we have the elementary result that 3DVAR, without averaging, cannot converge at fixed parameter choices:

Theorem 1.5 *Under Assumption 1 in dimension one, if u_n is generated by 3DVAR, then*

$$\mathbb{E}[|u_n - u^\dagger|^2] \geq \gamma^2\mathcal{K}^2.$$

As the method cannot converge in dimension one, it has no hope of converging in higher dimensions. By time averaging,

$$\bar{u}_n = \frac{1}{n}\sum_{k=1}^{n} u_k = \frac{1}{n}u_n + \frac{n-1}{n}\bar{u}_{n-1}, \tag{1.14}$$

we can obtain convergence for all $\alpha > 0$:

Theorem 1.6 *Under Assumption 1, fix $t \in [0, \nu]$ and $\tau_{\mathrm{v}} \in [0, 1]$, and, having set these indices, assume that $\Sigma^{t+1-\tau_{\mathrm{v}}(1+\nu)}$ is trace class. Then*

$$\mathbb{E}[\|\bar{u}_n - u^\dagger\|_t^2] \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{s+t}{1+\nu}} \|z_0\|^2 + \frac{\gamma^2}{\alpha}\mathrm{Tr}(\Sigma^{t+1-\tau_{\mathrm{v}}(1+\nu)})\left(\frac{n}{\alpha}\right)^{-\tau_{\mathrm{v}}}$$

where z_0 is the solution to

$$\Sigma^{-\frac{1}{2}}(u_0 - u^\dagger) = (B^*B)^{\frac{s-1}{2(1+\nu)}} z_0 \tag{1.15}$$

and $B = A\Sigma^{\frac{1}{2}}$.

We will repeatedly make use of the operator

$$B = A\Sigma^{\frac{1}{2}} \tag{1.16}$$

throughout this work. The existence of z_0 in (1.15) is a consequence of Assumption 1 on the initial error and an equivalence of spaces result encapsulated in Proposition 2.3, given below.

The motivation for time averaging comes from two related problems. First, formally, (1.4) has the structure of an AR(1) process, [23]. Under typical assumptions, an AR(1) process will not converge to a fixed value, but instead, sample an invariant distribution. Consequently, the time average will converge to the mean, with respect to this invariant distribution. Another motivation comes from the stochastic root finding problem and the Robbins-Monro algorithm. In [20], Polyak and Juditsky proved that by time averaging the sequence of estimates generated by Robbins-Monro, the convergence rate could be improved. See, also, [13].

As a consequence of Theorem 1.7, we will have unconditional mean squared convergence of the iterate-averaged value, $\bar{u}_n$, provided:

- We study the problem in a sufficiently weak weighted space ($t > 0$) and/or have sufficiently smooth data ($s > 0$);
- Σ has a sufficiently well behaved spectrum, allowing $\tau_{\mathrm{v}} > 0$. Note that taking $\tau_{\mathrm{v}} = t/(1+\nu)$ will not require additional assumptions on Σ, but will require $t > 0$ for convergence.

We emphasize that iterate averaging is a *post-processing* step, requiring no modification of the underlying 3DVAR iteration.

We introduce a modified version of Assumption 2,

Assumption 2′

(1) Σ and A^*A simultaneously diagonalize against the set $\{\varphi_i\}$ with respective eigenvalues σ_i and a_i^2, and these eigenvalues satisfy

$$\sigma_i \asymp i^{-1-2\epsilon}, \quad a_i \asymp i^{-p}, \quad \epsilon > 0, \quad p > 0. \tag{1.17}$$

(2) For $\beta \geq 0$, the initial error, $u_0 - u^\dagger$, satisfies the condition

$$\sum_{i=1}^{\infty} i^{2\beta} |u_{0,i} - u_i^\dagger|^2 < \infty. \tag{1.18}$$

Condition (1.18) on the initial error will automatically be satisfied if u_0 and $u^\dagger$ are, separately, sufficiently smooth. The assumptions of (1.17) and (1.18) are equivalent to those of (1.6) and (1.7) under the identifications:

$$\nu(1 + 2\epsilon) = 2p, \quad s(1 + 2\epsilon) = 2\beta.$$

In contrast to Assumption 2, no upper bound on β is necessary.

Theorem 1.7 *Under Assumption 2′, and having fixed a choice of $\|\bullet\|_t$ norm with $t \geq 0$, assume $\tau_b, \tau_v \in [0, 1]$ satisfy*

$$\tau_b \leq \frac{t(1 + 2\epsilon) + 2\beta}{2(1 + 2\epsilon + 2p)} \equiv \bar{\tau}_b \tag{1.19a}$$

$$\tau_v < \frac{t(1 + 2\epsilon) + 2\epsilon}{1 + 2\epsilon + 2p} \equiv \bar{\tau}_v \tag{1.19b}$$

then,

$$\mathbb{E}[\|\bar{u}_n - u^\dagger\|_t^2] \lesssim \left(\tfrac{n}{\alpha}\right)^{-2\tau_b} + \frac{\gamma^2}{\alpha}\left(\tfrac{n}{\alpha}\right)^{-\tau_v}$$

While our results in both the general and diagonal case establish unconditional convergence for any choice of α for the iterate-averaged 3DVAR, in a practical setting, there may only be n iterates available. One might then ask how well iterate-averaged 3DVAR behaves if, at fixed n, we choose the optimal α, and how this would compare to the minimax solution of (1.12). Focusing on the diagonal case, for comparison, we have the following result for the minimax solution of (1.12):

Theorem 1.8 *Under Assumption 2′ with $u_0 = 0$ in (1.18), if (1.12) is solved using a spectral cutoff with regularization α in the $t \geq 0$ norm, then at the optimal value of $\alpha = \alpha_\star(n)$,*

$$\mathbb{E}[\|\bar{u}_{n,\alpha_\star(n)} - u^\dagger\|_t^2] \lesssim \begin{cases} n^{-\frac{t(1+2\epsilon)+2\beta}{1+2p+2\beta}} & 1 + 2p \neq t(1 + 2\epsilon) \\ n^{-1} \log n & 1 + 2p = t(1 + 2\epsilon) \end{cases}$$

This is consistent with (1.13) and the results in [16, 17]. Then, looking at the minimax solution of 3DVAR, we obtain for two particular regimes:

Corollary 1.9 *With the same assumptions as Theorem 1.7, first, assume* $\bar{\tau}_\mathrm{b}, \bar{\tau}_\mathrm{v} \leq 1$. *Taking* $\tau_\mathrm{b} = \bar{\tau}_\mathrm{b}$ *and* $\tau_\mathrm{v} = (1-\theta)\bar{\tau}_\mathrm{v}$ *for* $\theta \in (0, 1]$,

$$\mathbb{E}[\|\bar{u}_n - u^\dagger\|_t^2] \lesssim \theta^{-\frac{2\beta+t(1+2\epsilon)}{1+2p+2\beta+\theta[t(1+2\epsilon)+2\epsilon]}} n^{-\frac{t(1+2\epsilon)+2\beta}{1+2p+2\beta+\theta[t(1+2\epsilon)+2\epsilon]}}.$$

If, instead, $\bar{\tau}_\mathrm{b}, \bar{\tau}_\mathrm{v} > 1$, *then, taking* $\tau_\mathrm{b} = \tau_\mathrm{v} = 1$,

$$\mathbb{E}[\|\bar{u}_n - u^\dagger\|_t^2] \lesssim n^{-1}$$

Consequently:

- At $t = 0$, in the first case,

$$\mathbb{E}[\|\bar{u}_n - u^\dagger\|_t^2] \lesssim \theta^{-\frac{2\beta}{1+2p+2\beta+2\epsilon\theta}} n^{-\frac{2\beta}{1+2p+2\beta+2\epsilon\theta}}$$

  This is somewhat better than (1.11), as there is no logarithmic term, and the factor of 2ϵ has been replaced by $2\epsilon\theta$, which can be reduced by taking θ smaller. The prefactor will grow, but it is independent of n.
- In the first case, where $\bar{\tau}_\mathrm{b}, \bar{\tau}_\mathrm{v} \leq 1$, by taking θ sufficiently close to zero, we can get arbitrarily close to the optimal rate in (1.13).
- The first case can be realized by taking t and β sufficiently small. The second case, where $\bar{\tau}_\mathrm{b}, \bar{\tau}_\mathrm{v} > 1$, is accessible by taking t large enough.
- There are two other cases to consider, $\bar{\tau}_\mathrm{b} \leq 1$, $\bar{\tau}_\mathrm{v} > 1$ and vice versa, but, for brevity we do not explore them here.

In contrast to iterate-averaged 3DVAR, there is no gain to iterate averaging for Kalman:

Theorem 1.10 *For the scalar Kalman filter, take* $C_0 = \frac{\gamma^2}{\alpha}\sigma > 0$. *Then the bias and variance of the iterate-averaged mean,* $\bar{m}_n$ *satisfy the inequalities*

$$\begin{aligned} |\mathbb{E}[\bar{m}_n] - u^\dagger| &\geq |\mathbb{E}[m_n] - u^\dagger|, \\ Var(\bar{m}_n) &\geq Var(m_n). \end{aligned}$$

Consequently, we do not further explore the impact of averaging upon the Kalman filter in this setting.

1.4 Outline

The structure of this paper is as follows. In Section 2 we review certain background results needed for our main results. Section 3 examines the scalar case, and it includes proofs of Theorems 1.5 and 1.10. We prove Theorems 1.6 and 1.7 in Section 4. Numerical examples are given in Section 5. We conclude with a brief discussion in Section 6.

2 Preliminary results

In this section, we establish some identities and estimates that will be crucial to proving our main results.

Much of our analysis relies on spectral calculus involving the following rational functions which are closely related to the Tikhonov-Phillips regularization $(\alpha + \lambda)^{-1}$:

$$r_{n,\alpha}(\lambda) = \left(\frac{\alpha}{\alpha+\lambda}\right)^n, \tag{2.1}$$

$$q_{n,\alpha}(\lambda) = \frac{1}{\lambda}\left\{1 - \left(\frac{\alpha}{\alpha+\lambda}\right)^n\right\} = \lambda^{-1}(1 - r_{n,\alpha}(\lambda)). \tag{2.2}$$

These are related by the identity

$$\sum_{k=1}^{m} r_{k,\alpha}(\lambda) = \alpha q_{m,\alpha}(\lambda). \tag{2.3}$$

The following estimates can be found in [10] and [19], particularly Section 2.2 of the latter reference:

Lemma 2.1 *For $\lambda \in [0, \Lambda]$ and $n \in \mathbb{N}$,*

$$0 < r_{n,\alpha}(\lambda) \leq \frac{\alpha}{\alpha + n\lambda} \leq 1,$$

$$\lambda^p r_{n,\alpha}(\lambda) \leq \begin{cases} \left(\frac{\alpha p}{n}\right)^p, & p \in [0, n], \\ \alpha^n \Lambda^{p-n}, & p > n. \end{cases}$$

Lemma 2.2 *For $\lambda \in [0, \Lambda]$, $n \in \mathbb{N}$,*

$$\lambda^p q_{n,\alpha}(\lambda) \leq \begin{cases} \left(\frac{n}{\alpha}\right)^{1-p}, & p \in [0, 1], \\ \Lambda^{p-1}, & p > 1, \end{cases}$$

$$\lambda^p q_{n,\alpha}(\lambda) \leq \lambda^{p-1}.$$

Next, we recall the following result on Hilbert scales,

Proposition 2.3 *There exists a constant $D > 1$, such that for $|\theta| \leq 1$,*

$$D^{-1}\|x\|_{\theta(1+\nu)} \leq \|(B^*B)^{\frac{\theta}{2}}x\| \leq D\|x\|_{\theta(1+\nu)}$$

and

$$\mathrm{Ran}\left((B^*B)^{\frac{\theta}{2}}\right) = \mathrm{Dom}\left(\Sigma_0^{-\frac{\theta(1+\nu)}{2}}\right).$$

This result, based on a duality argument, is proven in Lemma 4.1 of [10]. See, also, Section 8.4 of [6], particularly Corollary 8.22.

We also have a few useful identities for the filters which we state without proof.

Lemma 2.4 *For the Kalman filter, the mean and covariance operators and the Kalman gains satisfy the identities*

$$\begin{aligned} m_n &= \left(\gamma^2 n^{-1} C_0^{-1} + A^*A\right)^{-1} \left(A^*\bar{y}_n + \gamma^2 n^{-1} C_0^{-1} m_0\right) \\ C_n^{-1} &= C_{n-1}^{-1} + \gamma^{-2} A^*A = C_0^{-1} + \gamma^{-2} n A^*A \\ K_n &= (\gamma^2 C_{n-1}^{-1} + A^*A)^{-1} A^* = (\gamma^2 C_0^{-1} + nA^*A)^{-1} A^* = \gamma^{-2} C_n A^*. \end{aligned}$$

Lemma 2.5 *For 3DVAR,*

$$\bar{u}_n = \sum_{k=0}^{n-1} \frac{n-k}{n} (I - \mathcal{K}A)^k \mathcal{K} \bar{y}_{n-k} + \sum_{k=0}^{n-1} \frac{1}{n} (I - \mathcal{K}A)^k (I - \mathcal{K}A) u_0.$$

Corollary 2.6 *Letting* $v_n = u_n - u^\dagger$, $\bar{v}_n = \frac{1}{n}\sum_{k=1}^n v_k$,

$$\bar{v}_n = \sum_{k=0}^{n-1} \frac{n-k}{n} (I - \mathcal{K}A)^k \mathcal{K} \bar{\eta}_{n-k} + \sum_{k=0}^{n-1} \frac{1}{n} (I - \mathcal{K}A)^k (I - \mathcal{K}A) v_0.$$

Remark 2.7 As this is a linear problem, it will be sufficient to study the behavior of $\bar{v}_n$ to infer convergence of $\bar{u}_n$ to $u^\dagger$.

For the analysis of 3DVAR, the essential decomposition into bias and variance terms can be read off of Corollary 2.6. These can be expressed in the more useful forms using $q_{n,\alpha}$:

Lemma 2.8

$$\bar{I}_n^{bias} = \sum_{k=0}^{n-1} \frac{1}{n} (I - \mathcal{K}A)^k (I - \mathcal{K}A) v_0 = \frac{\alpha}{n} \Sigma^{\frac{1}{2}} q_{n,\alpha}(B^*B) \Sigma^{\frac{1}{2}} v_0, \tag{2.4}$$

$$\bar{I}_n^{var} = \sum_{k=0}^{n-1} \frac{n-k}{n} (I - \mathcal{K}A)^k \mathcal{K} \bar{\eta}_{n-k} = \frac{1}{n} \sum_{j=1}^{n} \Sigma^{\frac{1}{2}} q_{n-j+1,\alpha}(B^*B) B^* \eta_j. \tag{2.5}$$

Proof First, observe that

$$I - \mathcal{K}A = \Sigma^{\frac{1}{2}} \alpha (\alpha I + B^*B) \Sigma^{-\frac{1}{2}}.$$

Using this in (2.4) together with spectral calculus applied to positive self-adjoint compact operator B^*B, along with (2.3),

$$\begin{aligned} \bar{I}_n^{\text{bias}} &= \tfrac{1}{n} \textstyle\sum_{k=0}^{n-1} \Sigma^{1/2} \alpha^k (\alpha I + B^*B)^{-k+1} \Sigma_0^{-1/2} v_0 \\ &= \tfrac{1}{n} \textstyle\sum_{k=1}^{n} \Sigma^{1/2} r_{k,\alpha}(B^*B) \Sigma^{-1/2} v_0 = \tfrac{\alpha}{n} \Sigma^{\frac{1}{2}} q_{n,\alpha}(B^*B) \Sigma^{-\frac{1}{2}} v_0. \end{aligned}$$

Applying the same computations to (2.5), we have,

$$\begin{aligned} \bar{I}_n^{\text{var}} &= \textstyle\sum_{k=0}^{n-1} \frac{n-k}{n} \alpha^{-1} \Sigma_0^{\frac{1}{2}} r_{k+1,\alpha}(B^*B) B^* \bar{\eta}_{n-k} \\ &= \tfrac{1}{n} \textstyle\sum_{j=1}^{n} \left\{ \sum_{k=0}^{n-j} \alpha^{-1} \Sigma^{\frac{1}{2}} r_{k+1,\alpha}(B^*B) B^* \right\} \eta_j = \tfrac{1}{n} \textstyle\sum_{j=1}^{n} \Sigma^{\frac{1}{2}} q_{n-j+1,\alpha}(B^*B) B^* \eta_j. \end{aligned}$$

□

3 Analysis of the scalar problem

Before studying the general, infinite-dimensional case, it is instructive to consider the scalar problem, where $X = Y = \mathbb{R}$ and A, Σ, and $\mathcal{K}$ are now scalars. This setting will also allow us to establish the limitations of both 3DVAR and the Kalman filter.

3.1 3DVAR

First, we prove Theorem 1.5 which asserts that the 3DVAR iteration cannot converge in mean square:

Proof Since $y_n \sim \mathcal{N}(Au^\dagger, \gamma^2)$, we write $y_n = Au^\dagger + \eta_n$ for $\eta_n \sim \mathcal{N}(0, \gamma^2)$. By (1.4),

$$\begin{aligned} u_n - u^\dagger &= \mathcal{K}\eta_n + \mathcal{K}Au^\dagger + (1 - \mathcal{K}A)u_{n-1} - u^\dagger \\ &= \mathcal{K}\eta_n + (1 - \mathcal{K}A)(u_{n-1} - u^\dagger). \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{E}[|u_n - u^\dagger|^2] &= \mathbb{E}[|\mathcal{K}\eta_n|^2] + \mathbb{E}[|(1 - \mathcal{K}A)(u_{n-1} - u^\dagger)|^2] \\ &\geq \mathbb{E}[|\mathcal{K}\eta_n|^2] = \mathcal{K}^2\gamma^2. \end{aligned}$$

□

Next, studying the bias and variance of the time averaged problem, given by (2.4) and (2.5), we prove

Theorem 3.1 *For scalar time averaged 3DVAR, for* $\tau_b, \tau_v \in [0, 1]$

$$\mathbb{E}[|\bar{u}_n - u^\dagger|^2] \leq (A^2\Sigma)^{-2\tau_b}|v_0|^2 \left(\frac{n}{\alpha}\right)^{-2\tau_b} + \frac{\Sigma\gamma^2}{\alpha}(A^2\Sigma)^{-\tau_v}\left(\frac{n}{\alpha}\right)^{-\tau_v}.$$

Thus, we have unconditional convergence for any choice for $\alpha > 0$, something that we do not have for 3DVAR without any iterate averaging. The rate of convergence is greatest when $\tau_b \geq 1/2$ and $\tau_v = 1$.

To obtain the result, we make use of the bias-variance decomposition and expressions (2.4) and (2.5). In the scalar case, $B^*B = B^2 = \Sigma A^2$, so that

$$\left|\bar{I}_n^{\text{bias}}\right|^2 = \left(\frac{n}{\alpha}\right)^{-2} q_{n,\alpha}(\Sigma A^2)^2|v_0|^2. \tag{3.1}$$

Applying Lemma 2.2 to this expression, we immediately obtain

Proposition 3.2 *For* $0 \leq \tau_b \leq 1$,

$$\left|\bar{I}_n^{bias}\right|^2 \leq (A^2\Sigma)^{-2\tau_b}|v_0|^2 \left(\frac{n}{\alpha}\right)^{-2\tau_b}. \tag{3.2}$$

For the variance, we have the result

Proposition 3.3 *Let* $\tau_{\mathrm{v}} \in [0, 1]$,

$$\mathbb{E}[|\bar{I}_n^{var}|^2] \leq \frac{\Sigma\gamma^2}{\alpha}(A^2\Sigma)^{-\tau_{\mathrm{v}}}\left(\frac{n}{\alpha}\right)^{-\tau_{\mathrm{v}}}. \tag{3.3}$$

Proof For the scalar case of (2.5),using Lemma 2.2,

$$\begin{aligned}
\mathbb{E}[|\bar{I}_n^{\mathrm{var}}|^2] &= \tfrac{\gamma^2(A\Sigma)^2}{n^2}\textstyle\sum_{j=1}^n q_{j,\alpha}(A^2\Sigma)^2 \\
&= \tfrac{\gamma^2\Sigma}{n^2}(A^2\Sigma)^{1-(1+\tau_{\mathrm{v}})}\textstyle\sum_{j=1}^n\left[(A^2\Sigma)^{\frac{1+\tau_{\mathrm{v}}}{2}}q_{j,\alpha}(A^2\Sigma)\right]^2 \\
&\leq \tfrac{\Sigma\gamma^2}{n^2}(A^2\Sigma)^{-\tau_{\mathrm{v}}}\textstyle\sum_{j=1}^n\left(\frac{j}{\alpha}\right)^{2\left(1-\frac{1+\tau_{\mathrm{v}}}{2}\right)} \\
&\leq \tfrac{\Sigma\gamma^2(A^2\Sigma)^{-\tau_{\mathrm{v}}}}{n^2}n\left(\tfrac{n}{\alpha}\right)^{1-\tau_{\mathrm{v}}} = \tfrac{\Sigma\gamma^2}{\alpha}(A^2\Sigma)^{-\tau_{\mathrm{v}}}\left(\tfrac{n}{\alpha}\right)^{-\tau_{\mathrm{v}}}.
\end{aligned}$$

□

Proof of Theorem 3.1 The result then follows immediately by combining the two preceding propositions. □

3.2 Kalman filter

Here, we prove Theorem 1.10, showing there is no improvement in mean squared convergence of Kalman under iterate averaging.

Proof Using Lemma 2.4, for the k-the estimate of the mean,

$$\begin{aligned}
m_k &= \left(\tfrac{\alpha}{\Sigma k} + a^2\right)^{-1}\left(A\bar{y}_k + \tfrac{\alpha}{\Sigma k}m_0\right) \\
&= \left(\tfrac{\alpha}{\Sigma k} + A^2\right)^{-1}\left(A^2u^\dagger + A\bar{\eta}_k + \tfrac{\alpha}{\Sigma k}m_0\right) \\
&= \left(1 + \tfrac{\alpha}{A^2\Sigma k}\right)^{-1}u^\dagger + \left(1 + \tfrac{A^2\Sigma k}{\alpha}\right)^{-1}m_0 + \left(A + \tfrac{\alpha}{A\Sigma k}\right)^{-1}\bar{\eta}_k.
\end{aligned}$$

and without averaging,

$$\begin{aligned}
\mathbb{E}[m_n] - u^\dagger &= \left(1 + \frac{A^2\Sigma n}{\alpha}\right)^{-1}(m_0 - u^\dagger), \\
\mathrm{Var}(m_n) &= \left(A + \frac{\alpha}{A\Sigma n}\right)^{-2}\frac{\gamma^2}{n}.
\end{aligned}$$

Then, with averaging, for the bias,

$$\mathbb{E}[\bar{m}_n] - u^\dagger = \frac{1}{n}\sum_{k=1}^n\left(1 + \frac{A^2\Sigma k}{\alpha}\right)^{-1}(m_0 - u^\dagger),$$

and

$$\begin{aligned}
|\mathbb{E}[\bar{m}_n] - u^\dagger|^2 &= \left|\tfrac{1}{n}\textstyle\sum_{k=1}^n\left(1 + \frac{A^2\Sigma k}{\alpha}\right)^{-1}\right|^2|m_0 - u^\dagger|^2 \\
&\geq \left|\tfrac{1}{n}\textstyle\sum_{k=1}^n\left(1 + \frac{A^2\Sigma n}{\alpha}\right)^{-1}\right|^2|m_0 - u^\dagger|^2 = |\mathbb{E}[m_n] - u^\dagger|^2.
\end{aligned}$$

For the variance, first note

$$\bar{m}_n - \mathbb{E}[\bar{m}_n] = \frac{1}{n}\sum_{k=1}^{n}\left(A+\frac{\alpha}{A\Sigma k}\right)^{-1}\bar{\eta}_k = \frac{1}{n}\sum_{k=1}^{n}\left(A+\frac{\alpha}{A\Sigma k}\right)^{-1}\left\{\sum_{j=1}^{k}\eta_j\right\} = \frac{1}{n}\sum_{j=1}^{n}\eta_j\left\{\sum_{k=j}^{n}\left(A+\frac{\alpha}{A\Sigma k}\right)^{-1}\right\}.$$

Then, by dropping all but the $k = n$-th term in the inner sum,

$$\mathrm{Var}(\bar{m}_n) = \frac{1}{n^2}\sum_{j=1}^{n}\gamma^2\left\{\sum_{k=j}^{n}\left(A+\frac{\alpha}{A\Sigma k}\right)^{-1}\right\}^2 \geq \frac{1}{n^2}\sum_{j=1}^{n}\gamma^2\left(A+\frac{\alpha}{A\Sigma n}\right)^{-2} = \mathrm{Var}(m_n)$$

□

4 Analysis of the infinite-dimensional problem

We return to the bias and variance of 3DVAR in the general, potentially infinite-dimensional, setting and obtain estimates on the terms. We prove the general case in Section 4.1, and then the diagonal case in 4.2. Our minimax results are proven in Section 4.3.

4.1 General case

Here, we prove Theorem 1.6 by first establishing results on the bias and variance.

Proposition 4.1 *Under Assumption 1, with $t \in [0, \nu]$,*

$$\|\bar{I}_n^{bias}\|_t^2 \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{s+t}{1+\nu}}\|z_0\|^2$$

where z_0 solves (1.15).

The fastest possible decay available for the squared bias in Proposition 4.1 is $O(n^{-2})$ when $s = \nu + 2$ and $t = \nu$.

Proof We make use of bias term from Lemma 2.8, allowing us to write

$$\|\bar{I}_n^{\mathrm{bias}}\|_t^2 = \left\|\frac{\alpha}{n}\Sigma^{\frac{t+1}{2}}q_{n,\alpha}(B^*B)\Sigma^{-\frac{1}{2}}v_0\right\|^2.$$

Next, we make use of (1.6) and argue as in the Appendix of [10], applying Proposition 2.3. Since, by assumption, $v_0 \in \mathrm{Dom}(\Sigma^{-\frac{s}{2}})$, $\Sigma^{-\frac{1}{2}}v_0 \in \mathrm{Dom}(\Sigma^{-\frac{s-1}{2}})$. Then taking $\theta = (s-1)/(1+\nu)$ in the proposition, $\Sigma^{-\frac{1}{2}}v_0 \in \mathrm{Ran}((B^*B)^{\frac{s-1}{2(1+\nu)}})$ allows us to conclude the existence of z_0. Therefore,

$$\|\bar{I}_n^{\mathrm{bias}}\|_t^2 = \left\|\frac{\alpha}{n}\Sigma^{\frac{t+1}{2}}q_{n,\alpha}(B^*B)(B^*B)^{\frac{s-1}{2(1+\nu)}}z_0\right\|^2.$$

Next, using Proposition 2.3 again, now with $\theta = (1+t)/(1+\nu)$,

$$\begin{aligned}\|\bar{I}_n^{\text{bias}}\|_t^2 &\lesssim \left\| \tfrac{\alpha}{n}(B^*B)^{\frac{t+1}{2(1+\nu)}} q_{n,\alpha}(B^*B)(B^*B)^{\frac{s-1}{2(1+\nu)}} z_0 \right\|^2 \\ &= \left\| \tfrac{\alpha}{n}(B^*B)^{\frac{s+t}{2(1+\nu)}} q_{n,\alpha}(B^*B) z_0 \right\|^2 \\ &\leq \left(\sup_{0\leq\lambda\leq\|B^*B\|} \left| \tfrac{\alpha}{n} \lambda^{\frac{s+t}{2(1+\nu)}} q_{n,\alpha}(\lambda) \right| \right)^2 \|z_0\|^2 \leq \left(\tfrac{n}{\alpha}\right)^{-\frac{s+t}{1+\nu}} \|z_0\|^2.\end{aligned}$$

The last inequality holds since, $s \leq \nu+2$ and $t \leq \nu$, so that $0 \leq s+t \leq s+\nu \leq 2\nu+2$ allowing for the application of Lemma 2.2. □

Proposition 4.2 *Under Assumption 1, for $t \geq 0$, $\tau_{\text{v}} \in [0,1]$, and for this choice of τ_{v} and t, assume $\Sigma^{(1+t)-\tau_{\text{v}}(1+\nu)}$ is trace class. Then*

$$\mathbb{E}[\|\bar{I}_n^{var}\|_t^2] \lesssim \frac{\gamma^2}{\alpha} Tr(\Sigma^{t+1-\tau_{\text{v}}(1+\nu)}) \left(\frac{n}{\alpha}\right)^{-\tau_{\text{v}}}.$$

Remark 4.3 The fastest possible decay in the variance will be $\text{O}(n^{-1})$ when $\tau_{\text{v}} = 1$ and t is sufficiently large such that $\Sigma^{t-\nu}$ is trace class. However, the bias term requires $t \leq \nu$. This requires the identity operator to be trace class which will not hold in infinite dimensions.

Proof of Proposition 4.2 We begin with (2.5) and using that for any bounded operator T and positive self-adjoint trace class operator C, $|\text{Tr}(CT)| \leq \|T\||\text{Tr}C|$,

$$\begin{aligned}\mathbb{E}[\|\bar{I}_n^{\text{var}}\|_t^2] &= \tfrac{1}{n^2}\textstyle\sum_{j=1}^n \mathbb{E}[\|\Sigma^{\frac{t+1}{2}} q_{n-j+1,\alpha}(B^*B) B^* \eta_j\|^2] \\ &= \tfrac{\gamma^2}{n^2}\textstyle\sum_{j=1}^n \text{Tr}\left(\Sigma^{\frac{t+1}{2}} q_{j,\alpha}(B^*B)(B^*B) q_{j,\alpha}(B^*B)\Sigma^{\frac{t+1}{2}}\right) \\ &= \tfrac{\gamma^2}{n^2}\textstyle\sum_{j=1}^n \text{Tr}\left(\Sigma^{t+1-\tau_{\text{v}}(1+\nu)}\left(\Sigma^{\tau_{\text{v}}\frac{1+\nu}{2}}(B^*B)^{\frac{1}{2}} q_{j,\alpha}(B^*B)(B^*B)\right)^2\right) \\ &\leq \tfrac{\gamma^2}{n^2}\textstyle\sum_{j=1}^n \|\Sigma^{\tau_{\text{v}}\frac{1+\nu}{2}}(B^*B)^{\frac{1}{2}} q_{j,\alpha}(B^*B)(B^*B)\|^2 \text{Tr}(\Sigma^{t+1-\tau_{\text{v}}(1+\nu)}).\end{aligned}$$

Using Proposition 2.3 with $\theta = \tau_{\text{v}}$ and Lemma 2.2,

$$\begin{aligned}\|\Sigma^{\tau_{\text{v}}\frac{1+\nu}{2}}(B^*B)^{\frac{1}{2}} q_{j,\alpha}(B^*B)(B^*B)\| &\lesssim \|(B^*B)^{\frac{1+\tau_{\text{v}}}{2}} q_{j,\alpha}(B^*B)\| \\ &\lesssim \sup_{\lambda\in[0,\|B^*B\|]} \lambda^{\frac{1+\tau_{\text{v}}}{2}} q_{j,\alpha}(\lambda) \lesssim \left(\frac{j}{\alpha}\right)^{1-\frac{1+\tau_{\text{v}}}{2}}\end{aligned}$$

Therefore,

$$\mathbb{E}[\|\bar{I}_n^{\text{var}}\|_t^2] \lesssim \tfrac{\gamma^2}{n^2}\text{Tr}(\Sigma^{t+1-\tau_{\text{v}}(1+\nu)}) \textstyle\sum_{j=1}^n \left(\frac{j}{\alpha}\right)^{1-\tau_{\text{v}}} \lesssim \frac{\gamma^2}{\alpha}\text{Tr}(\Sigma^{t+1-\tau_{\text{v}}(1+\nu)}) \left(\frac{n}{\alpha}\right)^{-\tau_{\text{v}}}$$

□

Proof of Theorem 1.6 The theorem immediately follows from the two preceding propositions. □

4.2 Simultaneous diagonalization

A sharper result is available under the simultaneous diagonalization Assumption 2′. For convenience, letting

$$\omega = \frac{1+2\epsilon}{1+2\epsilon+2p}, \tag{4.1}$$

we have the relationship

$$\sigma_i \asymp (\sigma_i a_i^2)^{\omega}. \tag{4.2}$$

Proposition 4.4 *Under Assumption 2′, let* $\tau_{\mathrm{b}} \in [0,1]$ *satisfy condition* (1.19a),

$$\left\| \bar{I}_n^{bias} \right\|_t^2 \lesssim \left(\frac{n}{\alpha}\right)^{-2\tau_{\mathrm{b}}}.$$

Proof We start with (2.4) and then use (4.2) and Lemma 2.2,

$$\begin{aligned}
\left\| \bar{I}_n^{\mathrm{bias}} \right\|_t^2 &= \sum_{i=1}^{\infty} \left\langle \tfrac{\alpha}{n} \Sigma^{\frac{t+1}{2}} q_{n,\alpha}(B^*B) \Sigma^{-\frac{1}{2}} v_0 \right\rangle \varphi_i{}^2 \\
&= \sum_{i=1}^{\infty} \left| \tfrac{\alpha}{n} \sigma_i^{\frac{t}{2}} q_{n,\alpha}(\sigma_i a_i^2) \right|^2 \left| v_{0,i} \right|^2 = \left(\tfrac{\alpha}{n}\right)^2 \sum_{i=1}^{\infty} \sigma_i^t q_{n,\alpha}(\sigma_i a_i^2)^2 \left| v_{0,i} \right|^2 \\
&\asymp \left(\tfrac{n}{\alpha}\right)^{-2} \sum_{i=1}^{\infty} (\sigma_i a_i^2)^{t\omega - 2\tau_{\mathrm{b}}} ((\sigma_i a_i^2)^{\tau_{\mathrm{b}}} q_{n,\alpha}(\sigma_i a_i^2))^2 \left| v_{0,i} \right|^2 \\
&\lesssim \left(\tfrac{n}{\alpha}\right)^{-2} \left(\tfrac{n}{\alpha}\right)^{2-2\tau_{\mathrm{b}}} \sum_{i=1}^{\infty} (\sigma_i a_i^2)^{t\omega - 2\tau_{\mathrm{b}}} \left| v_{0,i} \right|^2
\end{aligned}$$

Using (1.19a),

$$\begin{aligned}
\sum_{i=1}^{\infty} (\sigma_i a_i^2)^{t\omega - 2\tau_{\mathrm{b}}} \left| v_{0,i} \right|^2 &\asymp \sum_{i=1}^{\infty} i^{-(1+2\epsilon+2p)(t\omega - i\tau_{\mathrm{b}})} \left| v_{0,i} \right|^2 \\
&\asymp \sum_{i=1}^{\infty} i^{-(1+2\epsilon+2p)(t\omega - 2\tau_{\mathrm{b}}) - 2\beta} i^{2\beta} \left| v_{0,i} \right|^2 \\
&\lesssim \left(\sup_i i^{-(1+2\epsilon+2p)(t\omega - 2\tau_{\mathrm{b}}) - 2\beta} \right) \sum_{i=1}^{\infty} i^{2\beta} \left| v_{0,i} \right|^2 < \infty
\end{aligned}$$

we have the result. □

Comparing this to the general case, we again see that if the data is sufficiently smooth and/or we study the problem in a sufficiently smooth space (β and/or t large), we can again obtain $\mathrm{O}(n^{-2})$ convergence of the squared bias.

Proposition 4.5 *Under Assumption 2′, having fixed* t, *for* $\tau_{\mathrm{v}} \in [0,1]$ *satisfying* (1.19b),

$$\mathbb{E}\left[\left\| \bar{I}_n^{var} \right\|_t^2 \right] \lesssim \frac{\gamma^2}{\alpha} \left(\frac{n}{\alpha}\right)^{-\tau_{\mathrm{v}}}$$

Proof Using (2.5), we begin by writing

$$\begin{aligned}
\mathbb{E}\left[\left\| \bar{I}_n^{\mathrm{var}} \right\|_t^2 \right] &= \tfrac{1}{n^2} \sum_{j=1}^{n} \mathbb{E}\left[\left\| \Sigma^{\frac{t+1}{2}} q_{n-j+1,\alpha}(B^*B) B^* \eta_j \right\|^2 \right], \\
&= \tfrac{\gamma^2}{n^2} \sum_{j=1}^{n} \mathrm{Tr}\left(\Sigma^{\frac{t+1}{2}} q_{n-j+1,\alpha} B^* B (B^*B) q_{n-j+1,\alpha}(B^*B) \Sigma^{\frac{t+1}{2}} \right), \\
&= \tfrac{\gamma^2}{n^2} \sum_{j=1}^{n} \mathrm{Tr}\left(\Sigma^{t+1} B^* B q_{j,\alpha}(B^*B)^2 \right).
\end{aligned}$$

Using (2.2) on each term in the sum,

$$\mathrm{Tr}\left(\Sigma^{t+1}B^*Bq_{j,\alpha}(B^*B)^2\right) = \sum_{i=1}^{\infty}\sigma_i^{t+2}a_i^2 q_{j,\alpha}(\sigma_i a_i^2)^2.$$

Then, using (4.2) and Lemma 2.2

$$\begin{aligned}\sigma_i^{t+2}a_i^2 q_{j,\alpha}(\sigma_i a_i^2)^2 &\asymp \sigma_i^{t+1}((\sigma_i a_i^2)^{\frac{1}{2}}q_{j,\alpha}(\sigma_i a_i^2))^2\\ &\asymp (\sigma_i a_i^2)^{\omega(t+1)}((\sigma_i a_i^2)^{\frac{1}{2}}q_{j,\alpha}(\sigma_i a_i^2))^2\\ &\asymp (\sigma_i a_i^2)^{\omega(t+1)-\tau_{\mathrm{v}}}((\sigma_i a_i^2)^{(1+\tau_{\mathrm{v}})/2}q_{j,\alpha}(\sigma_i a_i^2))^2\\ &\lesssim (\sigma_i a_i^2)^{\omega(t+1)-\tau_{\mathrm{v}}}\left(\frac{j}{\alpha}\right)^{1-\tau_{\mathrm{v}}}\end{aligned}$$

Under Assumption 1.19b

$$\sum_{i=1}^{\infty}(\sigma_i a_i^2)^{\omega(t+1)-\tau_{\mathrm{v}}} \asymp \sum_{i=1}^{\infty} i^{-[(1+2\epsilon)(t+1)-\tau_{\mathrm{v}}(1+2\epsilon+2p)]} < \infty \tag{4.3}$$

Consequently,

$$\mathrm{Tr}\left(\Sigma^{t+1}(B^*B)q_{j,\alpha}(B^*B)^2\right) \lesssim \left(\frac{j}{\alpha}\right)^{1-\tau_{\mathrm{v}}},$$

and

$$\frac{\gamma^2}{n^2}\sum_{j=1}^{n}\mathrm{Tr}\left(\Sigma^{t+1}(B^*B)q_{j,\alpha}(B^*B)^2\right) \lesssim \frac{\gamma^2}{\alpha}\left(\frac{n}{\alpha}\right)^{-\tau_{\mathrm{v}}}$$

□

In contrast to the non-diagonal case, if the problem is studied in a sufficiently weak sense (large enough t), one obtains $\mathrm{O}(n^{-1})$ convergence of the variance.

Proof of Theorem 1.7 This result immediately follows from the previous two propositions. □

4.3 Minimax analysis

Proof of Theorem 1.8 Recall the spectral cutoff regularization

$$g_\alpha(\lambda) = \lambda^{-1}1_{[\alpha,\infty)}(\lambda).$$

For a fixed α, the regularized solution of (1.12) is

$$\bar{u}_{n,\alpha} = g_\alpha(A^*A)A^*\bar{y}_n = g_\alpha(A^*A)A^*Au^\dagger + g_\alpha(A^*A)A^*\bar{\eta}_n.$$

This allows us to write the bias-variance decomposition of the error as

$$\mathbb{E}[\|\bar{u}_{n,\alpha} - u^\dagger\|_t^2] = \|(g_\alpha(A^*A)A^*A - I)u^\dagger\|_t^2 + \mathbb{E}[\|g_\alpha(A^*A)A^*\bar{\eta}_n\|_t^2]$$

For the bias term,

$$\|(g_\alpha(A^*A)A^*A - I)u^\dagger\|_t^2 = \sum_{i=1}^{\infty}\sigma_i^t(g_\alpha(a_i^2)a_i^2 - 1)^2|u_i^\dagger|^2.$$

Since $a_i^2 \asymp i^{-2p}$, all terms with $i \lesssim \alpha^{-\frac{1}{2p}}$ will vanish. This leaves us with

$$\begin{aligned}\|(g_\alpha(A^*A)A^*A - I)u^\dagger\|_t^2 &\lesssim \sum_{i=\lfloor \alpha^{-\frac{1}{2p}} \rfloor}^{\infty} \sigma_i^t |u_i^\dagger|^2 \\ &\lesssim \sum_{i=\lfloor \alpha^{-\frac{1}{2p}} \rfloor}^{\infty} i^{-t(1+2\epsilon)-2\beta} i^{2\beta} |u_i^\dagger|^2 \\ &\lesssim \alpha^{\frac{t(1+2\epsilon)+2\beta}{2p}} \sum_i i^{2\beta} |u_i^\dagger|^2.\end{aligned}$$

For the variance term, all terms with $i \gtrsim \alpha^{-\frac{1}{2p}}$ will vanish,

$$\begin{aligned}\mathbb{E}[\|g_\alpha(A^*A)A^*\eta\|_t^2] &= \frac{\gamma^2}{n} \sum_{i=1}^{\infty} \sigma_i^t g_\alpha(a_i^2)^2 a_i^2 \\ &\lesssim \frac{\gamma^2}{n} \sum_{i=1}^{\lceil \alpha^{-\frac{1}{2p}} \rceil} \sigma_i^t a_i^{-2} \lesssim \frac{\gamma^2}{n} \sum_{i=1}^{\lceil \alpha^{-\frac{1}{2p}} \rceil} i^{2p-t(1+2\epsilon)} \\ &\lesssim \begin{cases} \frac{\gamma^2}{n} \alpha^{-\frac{1+2p-t(1+2\epsilon)}{2p}} & 2p - t(1+2\epsilon) \neq -1 \\ \frac{\gamma^2}{n} \log \frac{1}{\alpha} & 2p - t(1+2\epsilon) = -1. \end{cases}\end{aligned}$$

Combining the two terms, we thus have,

$$\mathbb{E}[\|u - u^\dagger\|_t^2] \lesssim \begin{cases} \alpha^{\frac{t(1+2\epsilon)+2\beta}{2p}} + \frac{\gamma^2}{n} \alpha^{-\frac{1+2p-t(1+2\epsilon)}{2p}} & 2p - t(1+2\epsilon) \neq -1 \\ \alpha^{\frac{t(1+2\epsilon)+2\beta}{2p}} + \frac{\gamma^2}{n} \log \frac{1}{\alpha} & 2p - t(1+2\epsilon) = -1. \end{cases}$$

Optimizing over α yields the result. □

Proof of Corollary 1.9 Note that in the proof of Proposition 4.5, under our assumptions, in (4.3), with $\tau_v = (1-\theta)\bar{\tau}_v$

$$\begin{aligned}\sum_{i=1}^{\infty} (\sigma_i a_i^2)^{\omega(t+1)-\tau_v} &\asymp \sum_{i=1}^{\infty} i^{-[(1+2\epsilon)(t+1)-t(1+2\epsilon)(1-\theta)-2\epsilon(1-\theta)]} \\ &\asymp \sum_{i=1}^{\infty} i^{-1-\theta[t(1+2\epsilon)+2\epsilon]} \\ &\lesssim \int_1^{\infty} x^{-1-\theta[t(1+2\epsilon)+2\epsilon]} dx = \frac{1}{\theta[t(1+2\epsilon)+2\epsilon]}\end{aligned}$$

Consequently,

$$\mathbb{E}[\|\bar{u}_n - u^\dagger\|_t^2] \lesssim \left(\frac{n}{\alpha}\right)^{-\frac{t(1+2\epsilon)+2\beta}{1+2\epsilon+2p}} + \frac{1}{\theta}\frac{1}{\alpha}\left(\frac{n}{\alpha}\right)^{-(1-\theta)\frac{t(1+2\epsilon)+2\epsilon}{1+2\epsilon+2p}}$$

where the implicit constants in each term are independent of α, θ, and n. The optimally scaled α will be

$$\alpha = \theta^{-\frac{1+2p+2\epsilon}{1+2p+2\beta+\theta[t(1+2\epsilon)+2\epsilon]}} n^{\frac{2\beta-2\epsilon+\theta[t(1+2\epsilon)+2\epsilon]}{1+2p+2\beta+\theta[t(1+2\epsilon)+2\epsilon]}}$$

Substituting back in, we have our result. □

5 Numerical experiments

In this section we illustrate our results with some numerical experiments.

5.1 Scalar examples

As a simple scalar example, let $A = 1$, $\gamma = 0.1$, and $u^\dagger = 0.5$. For 3DVAR, take $u_0 = 0$, $\Sigma = 1$, and $\alpha = 1$, while for Kalman, take $m_0 = 0$ and $C_0 = 1$. Running 10^2 independent trials of each algorithm for 10^4 iterations, we obtain the results in Fig. 1. These simulations demonstrate our predictions from Theorems 1.5, Theorem 3.1, and Theorem 1.10, that 3DVAR can only converge with time averaging, while Kalman will not be improved by time averaging. The confidence bounds are computed using 10^4 bootstrap samples to produce 95% confidence intervals.

5.2 Simultaneous diagonalization example

Next, we consider the case of simultaneous diagonalization, working with functions in $L^2(0, 2\pi; \mathbb{R})$, and

$$A = (I - \tfrac{d^2}{dx^2})^{-1}, \quad \Sigma = A^2, \quad u^\dagger = 0. \tag{4.1}$$

The A operator is equipped with periodic boundary conditions, allowing us to easily work in Fourier space. As the problem is linear, we can separately consider the bias and the variance. In all examples below we discretize on $N = 2^{12}$ modes, and run for 10^4 iterations. This corresponds to $p = 2$ and $\epsilon = 1.5$ in Assumption $2'$.

For the bias, we choose, before truncation, as the initial condition

$$u_0 = \sum_{k=1}^{\infty} k^{-\frac{1}{2}-\beta-\delta} \cos(kx), \tag{4.2}$$

with $\beta = 1$ and $\delta = 0.01$. Consequently, this function satisfies (1.18) from Assumption $2'$. The perturbation δ is introduced so that we can best see the sharpness of our rates. Running the truncated and discretized problem, we obtain the results shown in Fig. 2 for the norms $t = 0, 0.5, 1, 2$. As the plots show, we are in good agreement with the maximal rate predicted by Theorem 1.7.

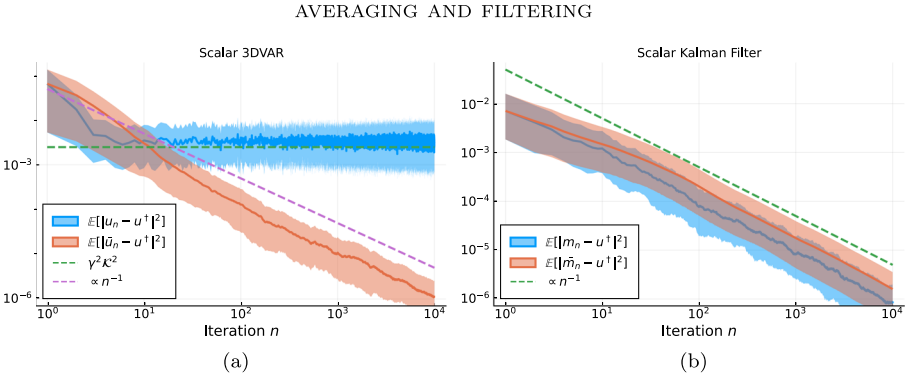

Fig. 1 Scalar results for 3DVAR and the Kalman filter. These results are consistent with Theorems 1.5, 1.10, and 3.1; 3DVAR will not converge without time averaging while Kalman will not improve from time averaging. Shaded regions reflect 95% confidence intervals at each n

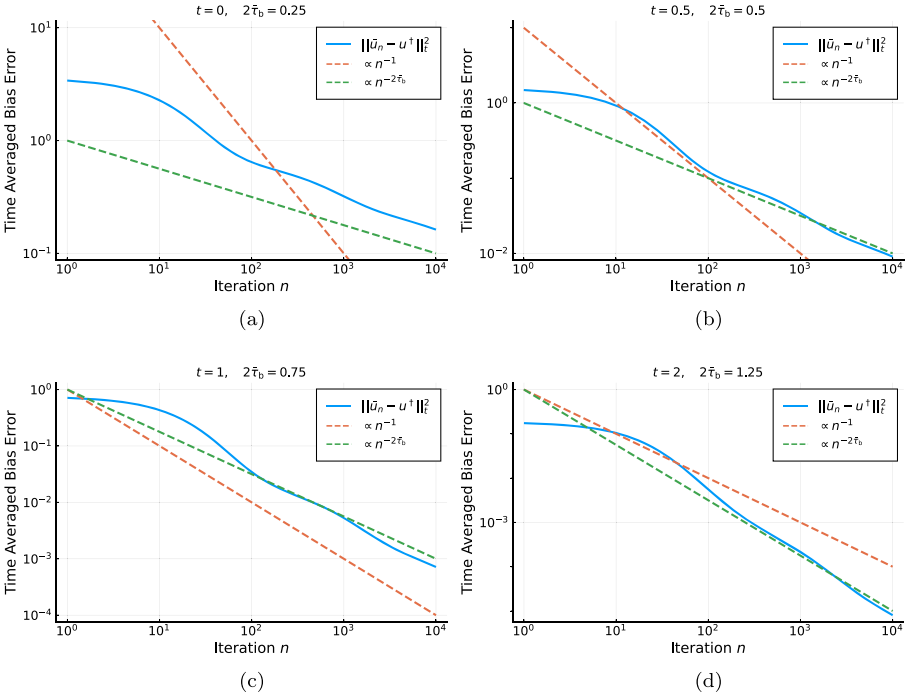

Fig. 2 Decay of the squared bias in our simultaneously diagonalized test problem for different t-norms. All are in good agreement with the rates predicted by Theorem 1.7. The constant $\bar{\tau}_v$ reflects the greatest possible decay rate from (1.19a)

For the variance, taking $u_0 = 0$, we run 10^2 independent trials of the problem, and then use bootstrapping to estimate 95% confidence intervals. The results, shown in Fig. 3, again show good agreement with the maximal rate predicted by Theorem 1.7.

6 Discussion

In this work we have examined the impact of iterate averaging upon the Kalman filter and 3DVAR as tools for solving a statistical inverse problem. We have found that this modest post-processing step ensures that the simpler algorithm, 3DVAR, will converge, unconditionally with respect to α, in mean square as the number of iterations $n \to \infty$. In contrast, there is no performance gain when this averaging is applied to the Kalman filter.

Our simulations suggest that our rates, at least in the diagonal case, may be sharp. For the diagonal case, we should expect to see something slower than the Monte Carlo rate of convergence, $O(n^{-1})$ unless working in a sufficiently weak space (large t). In the general case, it would seem that for the infinite-dimensional problem, we will never be able to achieve $O(n^{-1})$ convergence for the reasons outlined in Remark

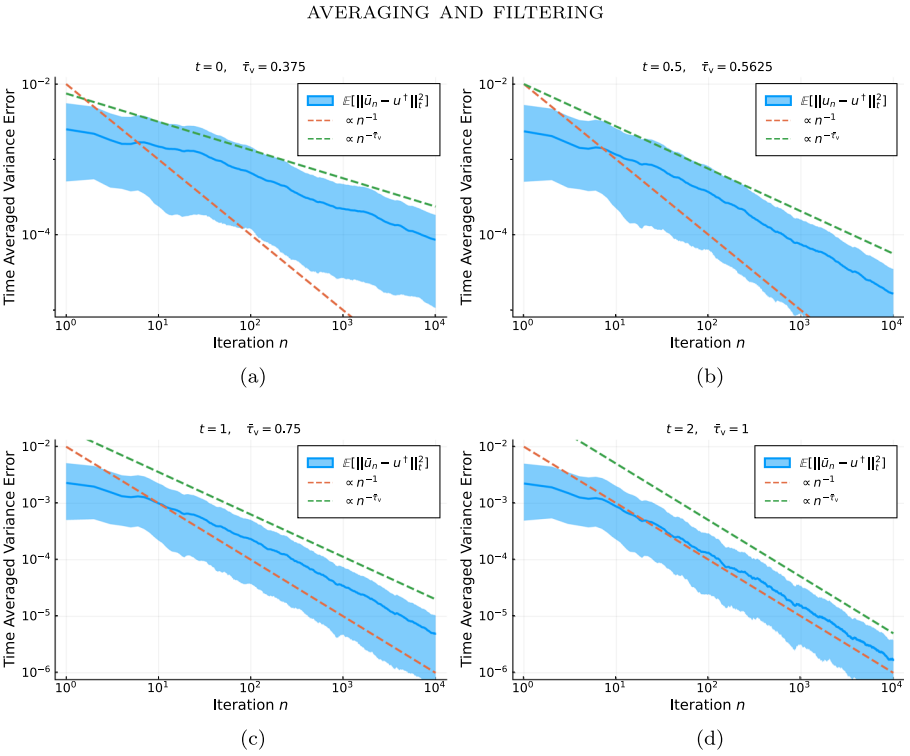

Fig. 3 Decay of the mean squared variance term in our simultaneously diagonalized test problem for different t-norms. All are in good agreement with the rates predicted by Theorem 1.7. Shaded regions reflect 95% confidence intervals at each n. The constant $\bar{\tau}_v$ reflects the greatest possible decay rate from (1.19b)

4.3; the operator Σ^{t-v} would need to be trace class, but $t \leq v$ for the bias to converge. The sharpness of the result in the non-diagonalizable case remains to be established. There is also potential for the extension of this work to the analogous continuous in time problem (1.5) studied in [15].

In actual applications, the problem will always be finite dimensional, making $O(n^{-1})$ achievable. In a spectral Galerkin formulation, truncating to N modes, and, $\mathrm{Tr}\Sigma_N^{t-v}$, will always be finite, though the constant may be large. Hence, we should expect to see $O(n^{-1})$ convergence, for sufficiently large n and a sufficiently severe dimensional truncation.

Acknowledgements The authors thank A.M. Stuart for suggesting an investigation of this problem. The content of this work originally appeared in [11] as a part of F.G. Jones's PhD dissertation. Work reported here was run on hardware supported by Drexel's University Research Computing Facility.

Funding This work was supported by US National Science Foundation Grant DMS-1818716.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Cavalier, L.: Nonparametric statistical inverse problems. Inverse Probl. **24**(3), 034004 (2008)
2. Da Prato, G., Zabczyk, J.: Stochastic equations in infinite dimensions. Cambridge University Press (2014)
3. Ding, L., Lu, S., Cheng, J.: Weak-norm posterior contraction rate of the 4dvar method for linear severely ill-posed problems. J. Complex. **46**, 1–18 (2018)
4. Ghanem, R., Higdon, D., Owhadi, H. (eds.): Handbook of Uncertainty Quantification. Springer, Cham (2017)
5. Ghosal, S., van der Vaart, A.: Fundamentals of nonparametric bayesian. inference Cambridge University Press (2017)
6. Heinz, W.E., Hanke, M., Neubauer, A.: Regularization of inverse problems. Kluwer Academic Publishers (2000)
7. Humpherys, J., Redd, P., West, J.: A fresh look at the Kalman filter. SIAM Rev. **54**(4), 801–823 (2012)
8. Iglesias, M.A.: A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. Inverse Probl. **32**, 025002 (2016)
9. Iglesias, M.A., Law, K.J., Stuart, A.M.: Ensemble kalman methods for inverse problems. Inverse Probl. **29**(4), 045001 (2013)
10. Iglesias, M.A., Lin, K., Lu, S., Stuart, A.M.: Filter based methods for statistical linear inverse problems. Commun. Math. Sci. **15**(7), 1867–1896 (2017)
11. Jones, F.G.E.: High and infinite-dimensional filtering methods. PhD thesis, Drexel University (2020)
12. Knapik, B.T., van der Vaart, A.W., van Zanten, J.H., et al.: Bayesian inverse problems with gaussian priors. Ann. Stat. **39**(5), 2626–2657 (2011)
13. Kushner, H.J., Yin, G.: Stochastic Approximation and Recursive Algorithms and Applications. Springer, New York (2003)
14. Law, K., Stuart, A., Zygalakis, K.: Data assimilation: a Mathematical Introduction. Springer International Publishing (2015)
15. Lu, S., Niu, P., Werner, F.: On the asymptotical regularization for linear inverse problems in presence of white noise. SIAM-ASA J. Uncertain. Quantif. **9**(1), 1–28 (2021)
16. Mair, B.A., Ruymgaart, F.H.: Statistical inverse estimation in hilbert scales. SIAM J. Appl. Math. **56**(5), 1424–1444 (1996)
17. Mathé, P., Pereverzev, S.V.: Optimal discretization of inverse problems in hilbert scales. regularization and self-regularization of projection methods. SIAM J. Numer. Anal. **38**(6), 1999–2021 (2001)
18. Øksendal, B.: Stochastic differential equations. Springer (2003)
19. Pereverzev, S., Lu, S.: Regularization theory for Ill-posed problems. De Gruyter (2013)
20. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. **30**(4), 838–855 (1992)
21. Schillings, C., Stuart, A.: Convergence analysis of ensemble Kalman inversion: the linear, noisy case. Appl. Anal., pp. 1–17 (2017)
22. Schillings, C., Stuart, A.M.: Analysis of the ensemble Kalman filter for inverse problems. SIAM J. Numer. Anal. **55**, 1264–1290 (2017)
23. Shumway, R.H., Stoffer, D.S.: Time series analysis and its applications. Springer (2011)
24. Stuart, A.M.: Inverse problems: A Bayesian perspective. Acta Numerica **19**, 451–559 (2010)
25. Sullivan, T.J.: Introduction to uncertainty quantification. Springer, vol. 63 (2015)
26. van Rooij, A.C., Ruymgaart, F.H.: Asymptotic minimax rates for abstract linear estimators. J. Stat. Plan. Inference **53**(3), 389–402 (1996)