Performance Guarantees for Score-Driven Filters

Simon Donker van Heel, Rutger-Jan Lange, Dick van Dijk, Bram van Os

February 23, 2024

Abstract

Tracking latent time-varying parameters in the presence of possible model misspecification is challenging, particularly when the true parameters exhibit large fluctuations and/or non-stationary dynamics. We derive performance guarantees for score-driven filters by presenting upper bounds for long-run root mean squared filtering errors. We distinguish between two classes of filters: *explicit* score-driven (ESD) and *implicit* score-driven (ISD). While the first class contains all score-driven filters in the literature, known variously as dynamic conditional score (DCS) or generalized autoregressive score (GAS) filters, the second class is essentially new. We relax conditions on the true parameter process considerably compared to recent work on error bounds for tracking latent time-varying parameters. These studies typically impose a limit on the true parameter variation, thereby excluding many realistic data generating processes, such as linear and Gaussian dynamics relevant to, for example, the Kalman filter. In contrast, we only necessitate a finite second moment for the (pseudo-)true parameter increments over time. Our theoretical analysis reveals, for the first time, that ESD filters require regularity conditions on the researcher-postulated logarithmic observation density. Specifically, Lipschitz continuity of the gradient, or equivalently, β-smoothness, is required to prevent the ESD filter from frequently 'overshooting' and, possibly, diverging to infinity. In contrast, ISD filters do not require this restrictive regularity condition. Indeed, our simulation studies across a wide variety of settings demonstrate that, when the true-parameter process is quite volatile, the ISD filter successfully tracks the true parameter even when the ESD filter diverges.

Keywords: Explicit and implicit-gradient methods; Error bounds; Misspecification

1 Introduction

Time-varying parameter models incorporate the idea that parameters may change over time. Numerous justifications exist for rejecting the assumption of constant parameters; we refer to Blasques et al. (2023) for a comprehensive overview. These models are broadly categorized into parameter-driven and observation-driven classes (Cox et al., 1981), although hybrid models combining both classes exist (Harvey, 1989). Parameter-driven models consider parameters as dynamic processes, each with their own source of randomness. Conversely, observation-driven models update parameters based on a function of observations, enabling maximum likelihood estimation (MLE) through the prediction error decomposition. An exploration of the differences between these classes is detailed in Koopman et al. (2016) through an extensive simulation study.

The tracking or filtering of unobserved time-varying parameters, or equivalently, states, is a fundamental problem in econometrics and statistics. Filtering concerns the real-time estimation of the states based on all past and current observations, assuming the model's static parameters are known. This task becomes more challenging in the presence of model misspecification, which is typically the case in practice, and when the true parameters feature large fluctuations and/or non-stationary dynamics.

Score-driven (SD) filters, variously known as dynamic conditional score (DCS; Harvey, 2013) or generalized autoregressive score (GAS; Creal et al., 2013) filters, find widespread applicability in both literature and practice, with over 300 articles available on the topic at `www.gasmodel.com`. These filters use the score of the conditional logarithmic likelihood function with respect to the parameter of interest to update state estimates. Their popularity stems from their generality, simplicity, and predictive capabilities, as showcased by, e.g. the recent Oxford encyclopedia entry by Artemova et al. (2022a,b).

In this article, we propose a new classification of score-driven filters into explicit score-driven (ESD) and implicit score-driven (ISD) filters. The explicit-implicit distinction, prevalent for decades in numerical analysis, see e.g. Ascher et al. (1995), is exemplified by methods related to SD filters such as stochastic gradient descent (SGD; Robbins and Monro, 1951). Remarkably, all score-driven filters in the existing literature, despite not explicitly using the term, are in fact *explicit* score-driven filters. With the exception of Lange et al. (2022), ISD filters have not been used in the time-varying parameter literature, and hence this class is essentially new.

ESD and ISD filters differ in their approach to estimating latent time-varying parameters. While ESD and ISD both employ linear first-order prediction steps, ISD employs an implicit update driven by the score evaluated in the updated parameter, thus appearing on both sides of the equation, a departure from ESD's explicit parameter-update where the score is evaluated in the predicted state. In the paper we show that standard quasi-Newton methods can solve the implicit update step under some regularity conditions. The implicit approach uses contemporaneous information, enhancing stability (Toulis et al., 2014) and allowing larger learning rates compared to explicit approaches (Toulis and Airoldi, 2017; Moulines and Bach, 2011; Ryu and Boyd, 2014).

We demonstrate that the ISD filter's parameter update step is the solution to an optimization problem maximizing the logarithmic observation density subject to a weighted quadratic penalty centered at the prediction. Replacing the log-density with its first-order approximation yields the ESD filter's parameter update step, characterizing the explicit version as a first-order approximation of the implicit filter; something pointed out by Lange et al. (2022), but not noticed in the extensive literature on score-driven models.

This article focuses on deriving performance guarantees for score-driven filters in tracking latent time-varying parameters, particularly in the context of possible model misspecification. The effectiveness of these filters is assessed through the derivation of upper bounds on long-run root mean squared filtering errors, where the sharpness of the bound offers a measure of accuracy in tracking latent states. A finite bound serves as a performance guarantee, preventing issues such as the difference between the true and filtered path diverging to infinity.

Our objective is to establish performance guarantees under minimal restrictions on the dynamics of the true parameter process. To enable tracking of the true parameter path over time, erratic variations must be ruled out. The minimal condition ensuring an upper bound on long-run root mean squared filtering errors is identified as a bounded second moment for any true parameter increment; this condition is considerably weaker than common assumptions in the existing literature. This expands the applicability of the performance guaranteed to a much broader range of data-generating processes; allowing e.g. unit root processes.

Our theoretical analysis reveals a new requirement for ESD filters to ensure a bounded filtering error over time. Lipschitz continuity of the gradient, or β-smoothness, of the researcher-postulated logarithmic observation density is deemed necessary to prevent 'overshooting' or divergence of the filter. In contrast, the ISD filter does not require this restrictive condition. While recent literature uses this condition as being sufficient for deriving certain optimality results (Gorgi et al., 2023), it fails to recognize that this condition may, in fact, be necessary. An illustration of a Poisson model with varying intensity in Section 3 highlights potential dramatic differences between implicit and explicit score-driven filters in tracking a scalar latent state when this condition is not met.

Filtering performance of explicit and implicit filters is explored in the choice of a penalty matrix. We derive that selecting a scalar multiple of the identity matrix minimizes the upper bounds on the long-run root mean squared error for both filters, ensuring a finite error bound for the implicit filter across all choices of (positive) penalty or learning rate, consistent with the related implicit SGD method (Toulis and Airoldi, 2017; Toulis et al., 2021). In contrast, the explicit filter requires a sufficiently low learning rate to avoid 'overshooting' or diverging. The sharpness of the error bound is strongly influenced by the curvature in the observation log-density, with constant curvature yielding the most precise bounds.

Naturally, these error bounds can be improved if we know more about the true parameter process. As a special case, we show that in a correctly specified local level model, the bound is tight, i.e. it can neither be surpassed nor improved further. The learning rate minimizing the bound is exactly the steady-state Kalman filter covariance, leading to optimality in the minimum mean squared error sense. Notably, both explicit and implicit filters can track unit root true parameter processes, but the identity prediction step is crucial for success.

A Monte Carlo study demonstrates that without Lipschitz continuity of the gradient, the ESD filter may frequently 'overshoot' or diverge to infinity, depending on the scaling of the score, while the ISD filter successfully tracks the true parameter. In conclusion, the ISD filter consistently outperforms ESD filters in terms of filtering RMSE across various scaling for the score, providing a stable and predictable performance even under large state variations or unit root dynamics.

2 Implicit and explicit score-driven filters

2.1 Problem setting

The $n \times 1$ observation $\boldsymbol{y}_t$ is drawn at times $t = 1, \ldots, T$ from a true (but typically unknown) conditional observation density $p^\dagger(\boldsymbol{y}_t|\boldsymbol{\theta}_t^\dagger, \boldsymbol{\psi}^\dagger, \mathcal{F}_{t-1})$. Here, $\boldsymbol{\theta}_t^\dagger$ a time-varying parameter vector taking its values in some parameter space $\boldsymbol{\Theta}^\dagger$, $\boldsymbol{\psi}^\dagger$ is a vector of static shape parameters, and $\mathcal{F}_{t-1}$ denotes the information set at time $t-1$. The inclusion of $\mathcal{F}_{t-1}$ allows the observation density to depend on exogenous variables and/or lags of $\boldsymbol{y}_t$. For readability, the dependence on $\boldsymbol{\psi}^\dagger$ and $\mathcal{F}_{t-1}$ will be suppressed. In the case of discrete observations, $p^\dagger(\boldsymbol{y}_t|\boldsymbol{\theta}_t^\dagger)$ is interpreted as a probability rather than a density. Here we do not specify the dynamics of the true process $\{\boldsymbol{\theta}_t^\dagger\}$; instead, we focus on the filtering method used by the researcher. We denote by $p(\boldsymbol{y}_t|\boldsymbol{\theta})$ the researcher-postulated density, which is typically misspecified, where $\boldsymbol{\theta} \in \mathbb{R}^d$ denotes a vector of parameters the researcher is interested in tracking over time. The assumption $\boldsymbol{\theta} \in \mathbb{R}^d$ is made for simplicity; if the parameter of interest has a natural range (e.g. being positive), standard link functions (such as the exponential function) may be employed. The hope is that the researcher-postulated density $p(\boldsymbol{y}|\boldsymbol{\theta})$ evaluated at the filtered parameter path, specified below, tracks the true density $p^\dagger(\boldsymbol{y}|\boldsymbol{\theta})$ evaluated at the true path $\{\boldsymbol{\theta}_t^\dagger\}$.

2.2 Filter specification

We consider *implicit* score-driven (ISD) and *explicit* score-driven (ESD) filters, where the nomenclature will shorty become apparent. The corresponding sequences of filtered states are denoted by $\{\boldsymbol{\theta}_{t|t}^{\mathrm{im}}\}$ and $\{\boldsymbol{\theta}_{t|t}^{\mathrm{ex}}\}$, respectively. Similarly, sequences of predictions are denoted by $\{\boldsymbol{\theta}_{t|t-1}^{\mathrm{im}}\}$ and $\{\boldsymbol{\theta}_{t|t-1}^{\mathrm{ex}}\}$. As is standard, filtered and predicted states reflect the researcher's estimates of the time-varying parameter using the contemporaneous information set $\mathcal{F}_t$ and lagged information set $\mathcal{F}_{t-1}$, respectively.

Given two initializations $\boldsymbol{\theta}^j_{0|0} \in \mathbb{R}^d$ with $j \in \{\text{im}, \text{ex}\}$, both filters employ a linear first-order prediction step for all $t = 1, \ldots, T$, i.e.

$$\text{prediction step:} \qquad \boldsymbol{\theta}^j_{t|t-1} \;=\; \boldsymbol{\omega} \,+\, \boldsymbol{\Phi}\, \boldsymbol{\theta}^j_{t-1|t-1}, \qquad j \in \{\text{im}, \text{ex}\}, \tag{1}$$

where $\boldsymbol{\omega}$ is the $d \times 1$ intercept, and $\boldsymbol{\Phi}$ is a $d \times d$ autoregressive matrix. While $\boldsymbol{\omega}$ and $\boldsymbol{\Phi}$ need not be identical for both filters (i.e. we allow $\boldsymbol{\omega}^j$ and $\boldsymbol{\Phi}^j$ with $j \in \{\text{im}, \text{ex}\}$), for readability their superscript may be suppressed. If the true process is (believed to be) a random walk, we may set $\boldsymbol{\omega} = \mathbf{0}$ and $\boldsymbol{\Phi} = \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ is the $d \times d$ identity. While in specific cases it may be possible to improve on this simple linear structure, in general we expect no immediate benefits as no (additional) information is revealed during the prediction step.

The crucial difference between both filters lies in their updating steps, which employ either implicit- or explicit-gradient methods, in both cases for all $t = 1, \ldots, T$:

$$\text{implicit-gradient update:} \qquad \boldsymbol{\theta}^{\text{im}}_{t|t} \;=\; \boldsymbol{\theta}^{\text{im}}_{t|t-1} \,+\, \boldsymbol{H}_t\, \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^{\text{im}}_{t|t}), \tag{2}$$

$$\text{explicit-gradient update:} \qquad \boldsymbol{\theta}^{\text{ex}}_{t|t} \;=\; \boldsymbol{\theta}^{\text{ex}}_{t|t-1} \,+\, \boldsymbol{H}_t\, \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^{\text{ex}}_{t|t-1}), \tag{3}$$

where $\boldsymbol{H}_t \in \mathbb{R}^{d \times d}$ is the $\mathcal{F}_{t-1}$-measurable learning-rate matrix, assumed symmetric and positive definite (i.e. $\boldsymbol{H}_t \succ \mathbf{O}$), $\nabla := \mathrm{d}/\,\mathrm{d}\boldsymbol{\theta}$ is the gradient operator acting on the second argument of $\ell(\boldsymbol{y}|\boldsymbol{\theta})$, and $\ell(\boldsymbol{y}|\boldsymbol{\theta}) := \log p(\boldsymbol{y}|\boldsymbol{\theta})$. Hence $\nabla \ell(\boldsymbol{y}|\boldsymbol{\theta})$ is the score, which explains part of the nomenclature. While the learning-rate matrix $\boldsymbol{H}_t$ need not be identical for both methods (i.e. we allow $\boldsymbol{H}^j_t$ with $j \in \{\text{im}, \text{ex}\}$), its superscript is suppressed when convenient. In the implicit update (2), the score on the right-hand side is evaluated at the point $\boldsymbol{\theta}^{\text{im}}_{t|t}$, which also appears on the left-hand side; this renders the method *implicit*, as the solution appears on both sides of the equation. In contrast, the explicit update (3) is immediately computable, since the score on the right-hand side is evaluated in the (explicitly known) prediction $\boldsymbol{\theta}^{\text{ex}}_{t|t-1}$, completing our motivation for the nomenclature. While ISD and ESD updates generally produce different outcomes, the example below illustrates a famous case in which both methods—albeit with different learning rates—yield identical filters.

Example 1 (Kalman filter is a special case of ISD ánd ESD filters) *Consider a correctly specified linear Gaussian state-space model, such that Kalman's (1960) filter applies; i.e. the observation is* $\boldsymbol{y}_t = \boldsymbol{d} + \boldsymbol{Z}\boldsymbol{\theta}^{\dagger}_t + \boldsymbol{\varepsilon}_t$*, where* $\boldsymbol{\varepsilon}_t \sim \text{i.i.d.}\, \text{N}(\boldsymbol{0}, \boldsymbol{R})$ *with* $\boldsymbol{R} \succ \mathbf{O}$*, while the latent state progresses as* $\boldsymbol{\theta}^{\dagger}_t = \boldsymbol{\omega} + \boldsymbol{\Phi}\boldsymbol{\theta}^{\dagger}_{t-1} + \boldsymbol{\eta}_t$*, where* $\boldsymbol{\eta}_t \sim \text{i.i.d.}\, \text{N}(\boldsymbol{0}, \boldsymbol{Q})$ *with* $\boldsymbol{Q} \succ \mathbf{O}$*. Denote Kalman's predicted and filtered states as* $\boldsymbol{\theta}^{\text{KF}}_{t|t-1}$ *and* $\boldsymbol{\theta}^{\text{KF}}_{t|t}$*, with corresponding covariance matrices* $\boldsymbol{P}^{\text{KF}}_{t|t-1}$ *and* $\boldsymbol{P}^{\text{KF}}_{t|t}$*. Kalman's predicted state has the linear first-order form* (47)*. Kalman's filtered state can be written (see Appendix X for details) as (i) ISD update with learning rate* $\boldsymbol{P}^{\text{KF}}_{t|t-1}$*, i.e.* $\boldsymbol{\theta}^{\text{KF}}_{t|t} = \boldsymbol{\theta}^{\text{KF}}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t-1}\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{KF}}_{t|t})$*, or as (ii) an ESD update with learning rate* $\boldsymbol{P}^{\text{KF}}_{t|t}$*, i.e.* $\boldsymbol{\theta}^{\text{KF}}_{t|t} = \boldsymbol{\theta}^{\text{KF}}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t}\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{KF}}_{t|t-1})$*. Hence, either the score or the learning rate is evaluated in the updated state. This "dual" representation of the Kalman filter has received little—if any—attention in the literature; most Kalman-filter extensions are based on its ESD representation (e.g. Fahrmeir, 1992, p. 504). While the learning rate* $\boldsymbol{P}^{\text{KF}}_{t|t}$ *for the ESD method should be* $\mathcal{F}_{t-1}$ *measurable, this is unproblematic as* $\boldsymbol{P}^{\text{KF}}_{t|t}$ *is known at time* $t-1$*, although it's not clear how easily generalizable this is. Clearly, the implicit learning rate exceeds the explicit one, as* $\boldsymbol{P}^{\text{KF}}_{t|t-1} \succ \boldsymbol{P}^{\text{KF}}_{t|t}$*. The equivalence of both methods (for different learning rates) is due to the linearity (in* $\boldsymbol{\theta}$*) of the score* $\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}) = \boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta})$*.*

2.3 Reformulation as optimization-based filters

Updating the parameter estimate in the direction of the gradient appears to be natural, as doing so would seem to improve the goodness of fit $\ell(\boldsymbol{y}_t|\boldsymbol{\theta})$—although in the case of explicit updates there are some caveats as discussed below. On the other hand, the step size governed by the learning rate $\boldsymbol{H}_t$ should not be excessively large; otherwise, the filtered path could become excessively volatile. To make explicit the trade-off between both competing aims (improving the goodness of fit, while maintaining stability over time), we reformulate both gradient-based updates in terms of an optimization framework.

Specifically, it is easy to check that the ISD update (2) is, in fact, the first-order condition corresponding to the following (multivariate) optimization problem:

$$\boldsymbol{\theta}_{t|t}^{\text{im}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmax}} \left\{ \ell\left(\boldsymbol{y}_t \mid \boldsymbol{\theta}\right) - \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\right\|_{\boldsymbol{P}_t}^2 \right\}. \tag{4}$$

Here, the penalty matrix $\boldsymbol{P}_t \in \mathbb{R}^{d\times d}$ is the inverse of the learning-rate matrix, i.e. $\boldsymbol{P}_t := \boldsymbol{H}_t^{-1} \succ \mathbf{O}$, while the penalty term $\|\cdot\|_{\boldsymbol{P}_t}^2 = (\cdot)'\boldsymbol{P}_t(\cdot)$ is a squared (weighted) Euclidean norm. The first-order condition associated with argmax (4) reads $\mathbf{0}_d = \nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - \boldsymbol{P}_t(\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}})$, which can be rearranged to yield the implicit update (2). As optimization (4) clarifies, the implicit update maximizes the logarithmic observation density $\ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ subject to a weighted quadratic penalty centered at the prediction; for this reason, optimization (4) is also known as a "proximal" method (reference).

The relation between both the ISD and ESD updates becomes apparent when in optimization problem (4) we linearly approximate (by a Taylor expansion) the logarithmic observation density around the prediction, which yields the explicit parameter update:

$$\boldsymbol{\theta}_{t|t}^{\text{ex}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmax}} \left\{ \underbrace{\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}^{\text{ex}}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_{t|t-1}^{\text{ex}})'\,\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}^{\text{ex}})}_{\approx\ell(\boldsymbol{y}_t|\boldsymbol{\theta})} - \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t|t-1}^{\text{ex}}\right\|_{\boldsymbol{P}_t}^2 \right\}. \tag{5}$$

The penalty matrix is again the inverse of the learning rate, i.e. $\boldsymbol{P}_t = \boldsymbol{H}_t^{-1} \succ \mathbf{O}$, which need not be identical for both methods. The objective function in optimization (5), in curly brackets, is linear-quadratic in $\boldsymbol{\theta}$; hence $\boldsymbol{\theta}_{t|t}^{\text{ex}}$ is analytically solveable. Indeed, the first-order condition is $\mathbf{0} = \nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{ex}}) - \boldsymbol{P}_t(\boldsymbol{\theta}_{t|t}^{\text{ex}} - \boldsymbol{\theta}_{t|t-1}^{\text{ex}})$, which can be rearranged to yield (3).

The downside of the linearized optimization (5) is that the intuitively appealing form (4) is lost. Moreover, while the full optimization (4) guarantees $\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) \geq \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{im}})$, i.e. the goodness of fit at each time step is improved, the same is not true for the linearized version (5). To explain, note that for the full optimization problem, the optimal value of the objective function (i.e. when evaluated at the argmax) must exceed the (suboptimal) value at any other point (e.g. at the predicted parameter). This fact yields $\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - 1/2\|\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\|_{\boldsymbol{P}_t}^2 \geq \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{im}})$. After rearrangement, we have $\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{im}}) \geq 1/2\|\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\|_{\boldsymbol{P}_t}^2 \geq 0$, which yields two desirable consequences: (i) the fit is improved at every time step, i.e. $\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{im}}) \geq 0$, and (ii) the stepsize is bounded, i.e. $\|\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\|_{\boldsymbol{P}_t} < \infty$, as long as the prediction is not arbitrarily bad and $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ is upper bounded, almost surely in $\boldsymbol{y}_t$. Hence the boundedness of the implicit update derives *not* from the boundedness of the gradient, but from the upper boundedness of the objective function itself, which provides a higher level of robustness.

In contrast, the solution (3) to the linearized update (5) may be prone to "overshooting"; i.e. unless the learning rate is very small, the undesirable situation $\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{ex}}) < \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{ex}})$ may regularly occur. In Section 4 we will find that for the explicit method to asymptotically achieve bounded filtering errors over time, we require that, almost surely in $\boldsymbol{y}_t$, the driving force $\boldsymbol{H}_t \nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{ex}})$ is Lipschitz in $\boldsymbol{\theta}_{t|t-1}^{\text{ex}}$. This additional condition, which is not needed for the implicit method, is required to avoid the explicit method from repeatedly overshooting and, possibly, diverging.

2.4 Computing the implicit-gradient update

Naturally, some assumptions (in Section 4) are required to ensure that the maximizer (4) exists. For example, the logarithmic observation density $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ being upper semi-continuous and concave, almost surely in $\boldsymbol{y}_t$, is sufficient but stronger than necessary. If $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ is also sufficiently smooth, then the unique global optimum can be found using standard quasi-Newton techniques (e.g. reference). For example, when $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ is concave and twice continuously differentiable, almost surely in $\boldsymbol{y}_t$, standard Newton-Raphson (NR, e.g. reference) iterates read

$$\boldsymbol{\theta}^{\text{im}}_{t|t} \;\leftarrow\; \boldsymbol{\theta}^{\text{im}}_{t|t} \,+\, \left[\boldsymbol{P}_t - \nabla^2 \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t})\right]^{-1} \left[\nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t}) \,-\, \boldsymbol{P}_t\,(\boldsymbol{\theta}^{\text{im}}_{t|t} - \boldsymbol{\theta}^{\text{im}}_{t|t-1})\,\right], \tag{6}$$

where $\nabla^2 := \nabla\nabla' = (\mathrm{d}/\,\mathrm{d}\boldsymbol{\theta})(\mathrm{d}/\,\mathrm{d}\boldsymbol{\theta})'$ denotes the Hessian operator. The algorithm may be initialized with $\boldsymbol{\theta}^{\text{im}}_{t|t} \leftarrow \boldsymbol{\theta}^{\text{im}}_{t|t-1}$. For high-dimensional problems, it may be beneficial to employ an algorithm that avoids large-matrix inversions, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (reference).

When computational efficiency is critical, algorithm (6) may be terminated after a single NR iteration, in which case the output (after one iteration) reads $\boldsymbol{\theta}^{\text{im}}_{t|t-1} + [\boldsymbol{P}_t - \nabla^2 \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t-1})]^{-1} \nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t-1})$. This "1NR" version is similar to the explicit update (3) in being computationally inexpensive; however, it is based on a quadratic (rather than linear) approximation of $\ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ around the prediction, which, as illustrated in Section 3, is advantageous when $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ exhibits strong curvature. On the other hand, additional iterations typically provide additional precision; hence, depending on the available computer power, researchers may decide to execute more or fewer iterations of algorithm (6). In our simulation studies, ~ 5 iterations typically provide an excellent level of accuracy.

2.5 Differences with literature on (explicit) score-driven filters

Here we compare our approach above with the well-known class of generalized autoregressive score (GAS; Creal et al., 2013) or dynamic conditional score (DCS; Harvey, 2013) filters. These filters have become collectively known as *score-driven filters* (e.g. Artemova et al., 2022a) and have found wide applicability; e.g. more than 300 articles are available on `www.gasmodel.com`. It turns out that this model class is nested within the framework presented above; more precisely, all score-driven filters in this literature are *explicit* score-driven filters, even as the word "explicit" is not typically used there. To demonstrate the equivalence, we substitute the explicit update (3) into the explicit prediction (47) to obtain the following (explicitly computable) prediction-to-prediction recursion: $\boldsymbol{\theta}^{\text{ex}}_{t+1|t} = \boldsymbol{\omega} + \boldsymbol{\Phi}\boldsymbol{\theta}^{\text{ex}}_{t|t-1} + \boldsymbol{\Phi}\boldsymbol{H}_t \nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{ex}}_{t|t-1})$. Up to reparameterization, this recursion is identical to that used in the (explicit) score-driven literature. To demonstrate this, we note that in this literature it is typical (e.g. Artemova et al., 2022a, p. 5) to take $\boldsymbol{H}_t = \boldsymbol{H}\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\text{ex}}_{t|t-1})^{-\zeta}$ with $\zeta \in \{0, 1/2, 1\}$, where $\boldsymbol{H} \succ \mathbf{O}$ is a static matrix, while the Fisher information matrix is $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) := \int p(\boldsymbol{y}|\boldsymbol{\theta}) \nabla \ell(\boldsymbol{y}|\boldsymbol{\theta}) \nabla' \ell(\boldsymbol{y}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{y} \succ \mathbf{O}$, which is positive definite under the usual identification assumptions. Interestingly, however, taking $\boldsymbol{H}_t = \boldsymbol{H}\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\text{ex}}_{t|t-1})^{-\zeta}$ guarantees neither symmetry nor positive definiteness of $\boldsymbol{H}_t$, even as imposing these properties would seem quite natural based on optimization (5).

Importantly for our purpose, we point out three important differences with the literature cited above. **First**, the literature on (explicit) score-driven filters has not considered splitting up the prediction-to-prediction recursion into distinct prediction and updating steps, as we have done here. We argue that differentiating between both steps is at once (i) familiar from state-space models and the Kalman filter (e.g. reference), (ii) conceptually useful, as new information is revealed during the updating (but not the prediction) step, and (iii) practically useful, as it may be beneficial to distinguish between predictions and nowcasts. Indeed, nowcasting studies using ESD filters are nonexistent as real-time updates are, in this literature, not defined. **Second**, the vast majority of the literature cited above assumes that the (explicit) score-driven filters are correctly specified in the sense that the filter, in fact, generated the data. Here we take the more realistic view that the data-generating process remains unknown, while we can only hope to show (in Section 4) that the ISD and ESD filters are reasonably accurate in tracking the (pseudo-)true time-varying parameter. **Third**, perhaps most fundamentally, the literature on (explicit) score-driven models has stopped short of recognizing that the ESD update (3) is identical to the solution of the linearized optimization problem (5). The lack of this connection being made has—arguably—prohibited researchers from considering the "full" optimization problem (4). With the exception of Lange et al. (2022), the implicit update (2) has, in the literature on time-varying parameters, not been employed. As our theory and simulations show, the ISD filter allows sharper error bounds to be derived; indeed, we will show that the ISD filter successfully tracks the true parameter even when the ESD filter diverges.

3 Illustration: Poisson data with varying intensity

o highlight the relevance of our theoretical results in the next section, here we illustrate the—in some cases, dramatic—differences between implicit and explicit score-driven filters in tracking a scalar latent state (i.e. $d = 1$) for a fundamental distribution.

True process: We consider Poisson-generated count data $y_t \in \mathbb{N}$ with a time-varying intensity $\lambda_t^\dagger = \exp(\theta_t^\dagger)$, where $\theta_t^\dagger \in \mathbb{R}$ is the time-varying parameter of interest, while the exponential link function ensures $\lambda_t^\dagger > 0$. We assume correct specification of the observation density, i.e. $p^\dagger(y|\lambda) = p(y|\lambda) = \lambda^y/y! \exp(-\lambda)$ for $y \in \mathbb{N}$ and $\lambda > 0$.

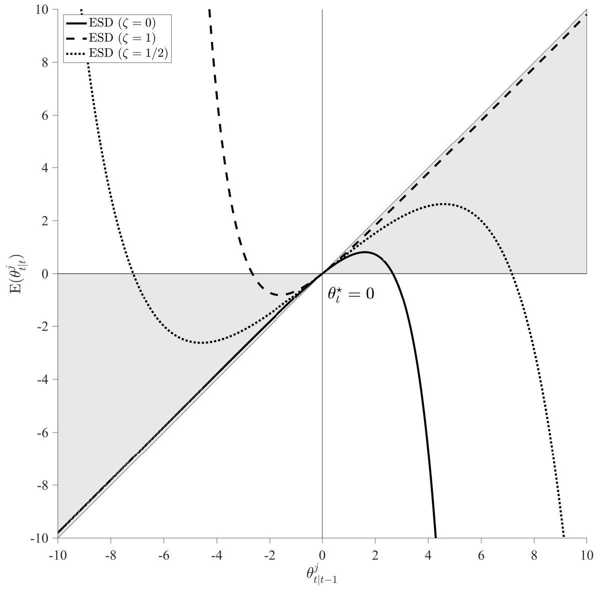


Figure 1: Expected update $\mathrm{E}(\theta^j_{t|t})$ when $\theta_t^\dagger = 0$ with learning rate $\eta^{\text{ex}} = 0.2$.

ISD filter: We implement the linear prediction (47) in combination with the implicit update (4). The filter is initialized using the true parameter value, i.e. $\theta^{\text{im}}_{0|0} = \theta_0^\dagger = 0$. The score is $\nabla\ell(y|\theta) = y - \exp(\theta)$ for $y \in \mathbb{N}$ and $\theta \in \mathbb{R}$, while the Hessian $\nabla^2\ell(y|\theta) = -\exp(\theta) < 0$ is strictly negative, such that $\theta \mapsto \ell(y|\theta)$ is strictly concave. Since the negative Hessian $-\nabla^2\ell(y|\theta) = \exp(\theta)$ does not depend on y, it equals the Fisher information quantity. We take a static learning rate $H_t^{\text{im}} = \eta^{\text{im}} > 0$, where η^{im} is a parameter to be estimated. The global maximizer (4) can be found using standard Newton-Raphson (NR) iterates (6).

ESD filter: We implement the linear prediction (47) in combination with the explicit update (3). As above, the filter is initialized using the true parameter value, i.e. $\theta^{\text{ex}}_{0|0} = \theta_0^\dagger = 0$. We follow the literature (e.g. Koopman et al., 2016) in taking the learning rate to be $H_t^{\text{ex}} = \eta^{\text{ex}} \exp(-\zeta\, \theta^{\text{ex}}_{t|t-1})$ with $\zeta \in \{0, 1/2, 1\}$, where $\eta^{\text{ex}} > 0$ is a static parameter to be estimated; hence, H_t^{ex} is time-varying unless $\zeta = 0$. The driving force in the explicit filter is $H_t^{\text{ex}}\nabla\ell(y_t|\theta^{\text{ex}}_{t|t-1}) = \eta^{\text{ex}} \exp(-\zeta\, \theta^{\text{ex}}_{t|t-1})(y - \exp(\theta^{\text{ex}}_{t|t-1}))$. This driving force contains exponential terms—except if $y = 0$ and $\zeta = 1$—and hence fails to be Lipschitz in the variable $\theta^{\text{ex}}_{t|t-1}$. Moreover, its average (over y) fails to be Lipschitz irrespective of $\zeta \in \{0, 1/2, 1\}$. To illustrate, suppose the true parameter is zero, i.e. $\theta^{\dagger}_t = 0$, such that the expected update equals

$$\mathrm{E}\left[\theta^{\text{ex}}_{t|t} \mid \mathcal{F}_{t-1}, \theta^{\dagger}_t = 0\right] = \theta^{\text{ex}}_{t|t-1} + \eta^{\text{ex}} \exp(-\zeta\, \theta^{\text{ex}}_{t|t-1})(1 - \exp(\theta^{\text{ex}}_{t|t-1})), \quad \zeta \in \{0, 1/2, 1\}, \tag{7}$$

where we have used $\mathrm{E}[y_t | \theta^{\dagger}_t = 0] = 1$.

Expected explicit-gradient update: Figure 1 shows the expected update (7) as a function of the prediction $\theta^{\text{ex}}_{t|t-1}$. Large *positive* prediction errors (corresponding to large values of $\theta^{\text{ex}}_{t|t-1} = \theta^{\text{ex}}_{t|t-1} - \theta^{\dagger}_t$) on average generate excessively large *negative* filtering errors for $\zeta = 0$ and $\zeta = 1/2$, as the corresponding curves exit the figure on the bottom right. Similarly, large *negative* prediction errors (corresponding to negative values of $\theta^{\text{ex}}_{t|t-1}$) on average generate exceedingly large *positive* filtering errors for $\zeta = 1/2$ and $\zeta = 1$, as the corresponding curves exit the figure on the top left. This behavior is illustrative of the tendency of explicit-gradient methods to overshoot; indeed, the magnitude of prediction errors can be arbitrarily magnified. Even as Koopman et al. (2016) advocate using $\zeta = 1/2$, we show here that the possibility of overshooting from both the left- and right-hand sides means that the filtered path may diverge, in alternating fashion, to infinity. In sum, for any fixed learning rate $\eta^{\text{ex}} > 0$, any true parameter $\theta^{\dagger}_t \in \mathbb{R}$ and any choice $\zeta \in \{0, 1/2, 1\}$, there exist (sufficiently inaccurate) predictions $\theta^{\text{ex}}_{t|t-1}$ for which the expected update $\mathrm{E}[\theta^{\text{ex}}_{t|t} | \theta^{\dagger}_t, \mathcal{F}_{t-1}]$ is exponentially divergent; for $\zeta = 1/2$, this may even imply a catastrophic runaway effect.

Expected implicit-gradient update: The expectation of the implicit-gradient update (2), also contained in Figure 1, is somewhat more involved; see Appendix X for the derivation. Figure 1 illustrates that, on average, the updated parameter $\theta^{j}_{t|t}$ is drawn closer to the true parameter $\theta^{\dagger}_t = 0$, i.e. $\mathrm{abs}(\mathrm{E}[\theta^{j}_{t|t} | \mathcal{F}_{t-1}, \theta^{\dagger}_t]) < \mathrm{abs}(\theta^{j}_{t|t-1})$, for implicit but *not* explicit updates. For the implicit update, the grey area in Figure 1 illustrates that $\mathrm{E}[\theta^{\mathrm{im}}_{t|t}|\mathcal{F}_{t-1},\theta^{\dagger}_{t}]$ has the same sign as $\theta^{\mathrm{im}}_{t|t-1}$, while being smaller in magnitude: on average, the prediction error is reduced in magnitude while its sign is unchanged.

4 Theory: Error bounds for score-driven filters

4.1 Assumptions

Our primary goal is to track the true density over time. Throughout, we consider the (root) mean-squared error ((R)MSE) relative to the pseudo-true parameter $\boldsymbol{\theta}^{\star}_{t}$ as the loss function of interest. The pseudo-true parameter $\boldsymbol{\theta}^{\star}_{t}$ is defined as the minimizer of the Kullback-Leibler divergence to the true density; when correctly specified, $\boldsymbol{\theta}^{\star}_{t}=\boldsymbol{\theta}^{\dagger}_{t}$.

Definition 1 (Pseudo-true parameter) *Consider a true distribution $p^{\dagger}(\boldsymbol{y}_t|\boldsymbol{\theta}^{\dagger}_{t})$ modeled by some postulated distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$. Then $\boldsymbol{\theta}^{\star}_{t}:=\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\ \int p^{\dagger}(\boldsymbol{y}|\boldsymbol{\theta}^{\dagger}_{t})\left[\ell^{\dagger}(\boldsymbol{y}|\boldsymbol{\theta}^{\dagger}_{t})-\ell(\boldsymbol{y}|\boldsymbol{\theta})\right]\mathrm{d}\boldsymbol{y}$ is the pseudo-true parameter, provided a unique solution exists, and $\ell^{\dagger}(\cdot|\theta^{\dagger}_{t}):=\log p^{\dagger}(\cdot|\boldsymbol{\theta}^{\dagger}_{t})$.*

To establish optimality guarantees, we require some regularity on both the true process and the postulated model. We first present mild assumptions on the evolution of the (pseudo-)true process to make tracking feasible. Afterwards, we discuss the regularity of our proposed observation density. Our core result on error bounds for score-driven filters concludes this section.

To be able to track the pseudo-true parameter path $\{\boldsymbol{\theta}^{\star}_{t}\}$ over time, it cannot be allowed to change haphazardly from one period to the next. Assumption 1a excludes this scenario for the misspecified case by imposing that the pseudo-true parameter increments $\{\boldsymbol{\theta}^{\star}_{t}-\boldsymbol{\theta}^{\star}_{t-1}\}$ have a finite second moment uniformly across time. This condition is needed to ensure that the RMSE loss can be computed; it is not related to our particular filtering setup.

Assumption 1b considers the correctly specified case, which is not a prerequisite for our main result but naturally allows for even stronger performance guarantees. Specifically, Assumption 1b states that the postulated density is correctly specified (such that $\boldsymbol{\theta}_t^\star = \boldsymbol{\theta}_t^\dagger$, for $t = 1, \ldots, T$) and that $\{\boldsymbol{\theta}_t^\dagger\}$ follows linear Gaussian dynamics with known coefficients, as is also assumed in deriving the Kalman filter. Assumption 1b therefore implies Assumption 1a.

Assumption 1 (Regularity of the (pseudo-)true process) *Consider a true distribution $p^\dagger(\boldsymbol{y}_t|\boldsymbol{\theta}_t^\dagger)$ modeled by some postulated distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$. Assume for $t = 1, \ldots, T$ that:*

(a) The pseudo-truth $\boldsymbol{\theta}_t^\star$ exists and is unique. In addition, the increments of the pseudo-true parameter have finite second (cross) moments. That is,

$$\mathrm{E}\left[\left(\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t-1}^\star\right)\left(\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t-1}^\star\right)'\right] \preceq \boldsymbol{Q}, \tag{8}$$

where $\boldsymbol{O}_d \preceq \boldsymbol{Q} \in \mathbb{R}^{d\times d}$ with $q^2 := \mathrm{tr}(\boldsymbol{Q}) < \infty$.

(b) The postulated density is correctly specified, i.e. $p(\boldsymbol{y}_t|\cdot) = p^\dagger(\boldsymbol{y}_t|\cdot)$ almost surely in $\boldsymbol{y}_t$, and the true parameter follows linear Gaussian state dynamics:

$$\boldsymbol{\theta}_t^\dagger = \boldsymbol{\omega}^\dagger + \boldsymbol{\Phi}^\dagger\boldsymbol{\theta}_{t-1}^\dagger + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \text{i.i.d.}\,\mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^\dagger), \tag{9}$$

with finite covariance matrix $\boldsymbol{O}_d \preceq \boldsymbol{Q}^\dagger \in \mathbb{R}^{d\times d}$ where $\sigma_\eta^2 := \mathrm{tr}(\boldsymbol{Q}^\dagger) < \infty$, intercept $\boldsymbol{\omega}^\dagger \in \mathbb{R}^d$ and auto-regressive matrix $\boldsymbol{\Phi}^\dagger \in \mathbb{R}^{d\times d}$ with spectral radius $\rho(\boldsymbol{\Phi}^\dagger) \leq 1$, which are assumed known.

Effectively, Assumption 1a implies the existence of a number of moments of the increments of the true process $\{\boldsymbol{\theta}_t^\dagger - \boldsymbol{\theta}_{t-1}^\dagger\}$, with the number depending on the form of postulated density and the severity of the misspecification. If the mapping from the true to the pseudo-true parameter is Lipschitz continuous—which trivially includes the correctly specified density case—then the boundedness of second moments of true-parameter increments is sufficient.

We emphasize that Assumption 1a is substantially weaker than common assumptions in the existing literature, which often consider bounded (pseudo-)true parameter variation (e.g. Wilson et al., 2018; Cao et al., 2019; Simonetto et al., 2020; Lanconelli and Lauria, 2023). These assumptions exclude most realistic data generating processes, such as the linear Gaussian dynamics in Assumption 1b, as would be relevant for e.g. the Kalman filter. In contrast, our theory is built on a minimal assumption needed for the loss function of interest to take finite value and does not presuppose correct specification; our core result will therefore be widely applicable. When correctly specified (i.e. Assumption 1b holds) these results may be further strengthened.

The score-driven filters of Section 2 connect the parameter update to the postulated observation density, the properties of which are therefore of paramount importance. Following standard practice in the gradient-based optimization literature (e.g. Boyd and Vandenberghe, 2004), Assumption 2 posits smoothness and concavity.

Assumption 2 (Regularity of the observation density) *Consider data $\boldsymbol{y}_t$ drawn from a true density $p^\dagger(\boldsymbol{y}_t|\boldsymbol{\theta}_t^\dagger)$ and modeled by some postulated distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta})$ that is once continuously differentiable in $\boldsymbol{\theta}$, almost surely in $\boldsymbol{y}_t$. Assume for $t = 1, \dots, T$ that either:*

(a) $\ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ is α-strongly concave in $\boldsymbol{\theta}$, $\alpha > 0$, almost surely in $\boldsymbol{y}_t$, or

(b) $\ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ is α-strongly concave and β-smooth in $\boldsymbol{\theta}$, where $0 < \alpha \leq \beta < \infty$, almost surely in $\boldsymbol{y}_t$.

Assumption 2a is sufficient to ensure that the score on average points in the direction of the pseudo-truth $\boldsymbol{\theta}_t^\star$, while its magnitude grows sufficiently fast as the prediction $\boldsymbol{\theta}_{t|t-1}^j$, $j \in \{\mathrm{im}, \mathrm{ex}\}$ is moved away from $\boldsymbol{\theta}_t^\star$. Furthermore, Assumption 2b additionally requires that the postulated log likelihood is β-smooth. This is equivalent to β-Lipschitz gradient continuity and limits the maximum growth of the score and/or the curvature of the postulated logarithmic observation density $\ell(\boldsymbol{y}_t|\cdot)$. Assumption 2b turns out to be a necessary ingredient to obtain a finite error bound for the ESD—but not the ISD—filter.

Assumption 3 (Bounded information) *Consider a true distribution $p^\dagger(\boldsymbol{y}_t|\boldsymbol{\theta}_t^\dagger)$ modeled by some postulated distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$. Assume for $t = 1, \ldots, T$ that*

$$\sigma_t^2 \;:=\; \int p^\dagger(\boldsymbol{y}|\boldsymbol{\theta}_t^\dagger) \|\nabla \ell(\boldsymbol{y}|\boldsymbol{\theta}_t^\star)\|^2 \, \mathrm{d}\boldsymbol{y} \;\leq\; \sigma_{\max}^2 \;<\; \infty. \tag{10}$$

While Assumptions 1–3 allow a rich combination of DGPs and postulated observation densities, we explicitly point out an important special case that is allowed in our setup but often ruled out in other literature.

Example 2 (Linear Gaussian state-space model) *Consider a correctly specified linear Gaussian state-space model, for which the observation equation reads $\boldsymbol{y}_t = \boldsymbol{d} + \boldsymbol{Z}\boldsymbol{\theta}_t^\dagger + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t \sim \text{i.i.d.}\,\mathrm{N}(\mathbf{0}, \boldsymbol{R})$ and $\boldsymbol{R} \succ \mathbf{O}$ is a positive-definite covariance matrix with finite trace $r = \mathrm{tr}(\boldsymbol{R}) < \infty$, while the state transition satisfies the linear Gaussian dynamics* (46) *in Assumption 1b. The negative Hessian of the logarithmic observation density reads $-\nabla^2 \ell(\boldsymbol{y}_t|\boldsymbol{\theta}) = \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}$; hence, Assumption 2b is satisfied with minimum and maximum curvature $\alpha = \lambda_{\min}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}) > 0$ and $\beta = \lambda_{\max}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}) < \infty$, respectively. Finally, Assumption 3 is satisfied as* $\mathrm{E}[\|\nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_t^\dagger)\|^2] = \mathrm{E}[\|\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}_t^\dagger)\|^2] = \mathrm{E}[\|\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{\varepsilon}_t\|^2] = \mathrm{E}[\mathrm{tr}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\boldsymbol{R}^{-1}\boldsymbol{Z})] = \mathrm{tr}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\,\mathrm{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t']\boldsymbol{R}^{-1}\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}) \leq$ *$d\lambda_{\max}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}) = d\beta < \infty$.*

4.2 Error bounds

First, let us introduce some notation. For a $d \times d$ matrix $\boldsymbol{A}$, we write $\|\boldsymbol{A}\|_2 := \sqrt{\lambda_{\max}(\boldsymbol{A}'\boldsymbol{A})}$ for the spectral norm. Here, $\lambda_{\max}(\boldsymbol{A})$ denotes the maximum eigenvalue of matrix $\boldsymbol{A}$. We consider the root mean-squared error of the filtered parameter relative to the pseudo-true parameter $\mathrm{RMSE}_{t|t} := \sqrt{\mathrm{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2]}$ as the loss function of interest. Furthermore, we denote by s^2 the maximum of the second moment of the true state over time, which may be non-finite $\mathrm{E}[\|\boldsymbol{\theta}_t^\star\|^2] \leq s^2$ for $t = 1, \ldots, T$. Let $\mu_{\min}$ and $\mu_{\max}$ represent the minimum and maximum eigenvalue of penalty matrix $\boldsymbol{P}_t$, respectively.

Here, we provide performance guarantees for ISD and ESD filters under possible model misspecification. More precisely, Theorem 1 presents upper bounds for long-run root mean squared filtering errors (RMSEs) under the very mild conditions on the (pseudo-)true parameter process given in Assumption 1a. The ISD and ESD filter's error bounds (12) and (14) are guaranteed under the sufficient contraction conditions (11) and (13) on the filter's specification and postulated distribution.

Theorem 1 (RMSE bounds under misspecification) *Let the postulated conditional density $p(\cdot|\cdot)$ be such that Assumptions 1a and 3 hold. Additionally, let the ISD filter satisfy Assumption 2a, and the ESD filter Assumption 2b.*

1) Consider the ISD *filter, i.e. prediction* (47) *and* implicit *update* (2). *Let*

$$\frac{\sqrt{\mu_{\max}}\,\|\boldsymbol{\Phi}\|_2}{\sqrt{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}}} < 1. \tag{11}$$

Then

$$\limsup_{t\to\infty} \mathrm{RMSE}_{t|t} \leq \frac{\frac{\sigma_{\max}}{\sqrt{\mu_{\min}}} + \sqrt{\mu_{\max}}\,[q + \|\boldsymbol{\omega}\| + \|\boldsymbol{\Phi} - \boldsymbol{I}_d\|_2\, s]}{\sqrt{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}} - \sqrt{\mu_{\max}}\|\boldsymbol{\Phi}\|_2}. \tag{12}$$

2) Consider the ESD *filter, i.e. prediction* (47) *and* explicit *update* (3). *Let*

$$\left(\sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} + \frac{\beta}{\mu_{\min}}\right)\|\boldsymbol{\Phi}\|_2 < 1. \tag{13}$$

Then

$$\limsup_{t\to\infty} \mathrm{RMSE}_{t|t} \leq \frac{\frac{\sigma_{\max}}{\sqrt{\mu_{\min}}} + \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}}\right)\left[q + \|\boldsymbol{\omega}\| + \|\boldsymbol{\Phi} - \boldsymbol{I}_d\|_2\, s\right]}{\sqrt{\mu_{\min}} - \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}}\right)\|\boldsymbol{\Phi}\|_2}. \qquad (14)$$

The upper bounds on the long-run RMSEs of both score-driven filtering classes are minimized for $\mu_{\min} = \mu_{\max}$, or equivalently, when the penalty matrix is a scalar multiple of the identity matrix $\boldsymbol{P}_t = \gamma \boldsymbol{I}_d$, which automatically holds in the univariate case. Hence, this choice for the penalty matrix is not only practically useful as less parameters need to be estimated, it is also theoretically supported. Moreover, the RMSE bounds and contraction conditions from Theorem 1 simplify considerably, as depicted in the Appendix.

Corollary 1.1 *Let the conditions of Theorem 1 hold, and let the penalty matrix be a scalar multiple of the identity matrix $\boldsymbol{P}_t = \gamma \boldsymbol{I}_d$. Then the contraction condition for the ISD filter reads $\frac{\gamma}{\gamma+\alpha}\|\boldsymbol{\Phi}\|_2 < 1$, such that* 1) *the RMSE bound is minimal,* 2) *the contraction condition is automatically satisfied for any non-explosive autoregressive matrix $\|\boldsymbol{\Phi}\|_2 \leq 1$.*

Corollary 1.1 implies that for this choice of penalty, the ISD filter can always guarantee an upper bound on the long-run RMSE. In contrast, the ESD filter cannot generally ensure bounded errors when the learning rate is a scalar multiple of the identity, as it depends on the penalty size γ and the minimum and maximum curvature in the observation log-density, denoted by α and β, respectively. However, the ESD RMSE bound is similarly minimized for $\boldsymbol{P}_t = \gamma \boldsymbol{I}$.

Consider a trivial filter that persistently estimates the latent state by its unconditional mean, disregarding any variations in the level. This filter achieves an upper bound on long-run RMSE for mean-reverting true states. However, if the true state tends to infinity, due to, for instance, unit root dynamics, this trivial filter cannot guarantee such an upper bound. In essence, tracking states with non-stationary dynamics poses greater difficulty for a filter. Corollary 1.2 ensures that even with unit root dynamics $\mathbf{\Phi}^\dagger = \boldsymbol{I}_d$ causing $s \to \infty$, ESD and ISD filters are capable of accurately tracking latent states in the long run.

Corollary 1.2 *Let the conditions of Theorem 1 hold. Both ISD and ESD filters can guarantee an upper bound on the long-run root mean squared filtering error for tracking latent states with unit root dynamics using an identity transformation ($\boldsymbol{\omega} = \mathbf{0}$ and $\mathbf{\Phi} = \boldsymbol{I}_d$) in the filter's prediction step* (47)*, provided that the relevant contraction condition is satisfied.*

The RMSE bounds in Theorem 1 are derived under minimal assumptions on the true parameter process, i.e. Assumption 1a. Consequently, these bounds can be sharpened with more available information about the true parameter process is available. In the derivation of the Kalman filter, it is assumed that the true parameter process is correctly specified with a linear and Gaussian state-transition equation. Theorem 2 presents the RMSE bounds and contraction conditions of the ISD and ESD filter under this assumption (1b).

Theorem 2 ((R)MSE bounds under correct specification) *Let the postulated conditional density $p(\cdot|\cdot)$ be such that Assumptions 1b and 3 hold. Additionally, let the ISD filter satisfy Assumption 2a, and the ESD filter Assumption 2b.*

1) Consider the ISD *filter, i.e. prediction* (47) *and* implicit *update* (2)*. Let*

$$\frac{\mu_{\max}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}} \|\mathbf{\Phi}\|_2^2 < 1, \tag{15}$$

Then

$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\frac{\sigma_{\max}^2}{\mu_{\min}} + \mu_{\max}\sigma_\eta^2}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}} - \mu_{\max}\|\mathbf{\Phi}\|_2}. \tag{16}$$

2) Consider the ESD *filter, i.e. prediction* (47) *and* explicit *update* (3)*. Let*

$$\left(\sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} + \frac{\delta}{\mu_{\min}}\right) \|\mathbf{\Phi}^\star\|_2 < 1 \tag{17}$$

Then

$$\limsup_{t\to\infty} \mathrm{RMSE}_{t|t} \leq \frac{\frac{\sigma_{\max}}{\sqrt{\mu_{\min}}} + \sigma_\eta\left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}}\right)}{\sqrt{\mu_{\min}} - \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}}\right)\|\boldsymbol{\Phi}^\star\|_2}. \tag{18}$$

The error bound's precision is strongly influenced by the curvature in the observation log-density, specifically, by α and β denoting the minimum and maximum curvature. Specifically, the bounds decrease with α and increase with β. Hence, a model with constant curvature in the observation log-density produces the most precise bounds. Example 3 illustrates that a correctly specified (Gaussian) local level model with constant curvature achieves a tight RMSE bound, in the sense that it can neither be surpassed nor improved further. This is derived by establishing that the learning rate minimizing this bound equals the inverse of the steady-state Kalman filter covariance for the local level model in Durbin and Koopman (2012). As the Kalman filter is optimal in the minimum mean squared error sense, the bound is tight.

Example 3 (MSE bounds ISD filter local level model) *Let the data be generated from a local level model:* $y_t = \theta_t^\star + \varepsilon_t, \theta_{t+1}^\star = \theta_t^\star + \eta_t, \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2), \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2)$*, where* ε_t *and* η_t *are mutually independent for all* t*, such that* $\boldsymbol{P} = \gamma$*,* $\sigma_{\max}^2 = \frac{1}{\sigma_\epsilon^2}$ *and* $\alpha = \frac{1}{\sigma_\epsilon^2}$*. Suppose we are correctly specified, such that Assumptions 1b, 2b, and 3 are satisfied, then*

$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\sigma_\varepsilon^2 + \sigma_\eta^2\sigma_\varepsilon^4\gamma^2}{2\gamma\sigma_\varepsilon^2 + 1}, \tag{19}$$

which is minimized for penalty

$$\hat{\gamma} = \frac{2}{\left(\sigma_\eta^2 + \sigma_\eta^2\sqrt{4\sigma_\varepsilon^2 + \sigma_\eta^2}\right)} =: \frac{1}{\bar{p}}, \tag{20}$$

where $\bar{p}$ *is exactly the steady-state Kalman filter covariance for the local level model.*

5 Simulation

This Monte Carlo study emphasizes the importance of the theoretical results in the preceding section. It illustrates the potential implications of employing an ESD filter without sufficient regularity in the researcher-postulated logarithmic observation density, specifically β-smoothness. Furthermore, we show that when the true process is volatile, the ISD filter successfully tracks the true parameter, even in cases where the ESD filter diverges.

We consider nine univariate data-generating processes (DGPs) from Koopman et al. (2016) with linear Gaussian state dynamics, thereby satisfying Assumption 1a. Table 1 from Lange (2020) depicts an overview of the observation densities, score and link functions, and other relevant quantities. The first seven DGPs are log-concave, while the Gaussian and Student-t copula (SCg and SCt) models are not. Here, we focus on the Poisson model with time-varying intensity from the Illustration 3, but similar results hold for the other eight models from Koopman et al. (2016) as shown in the Appendix.

Table 1: Overview of data-generating processes in simulation studies.

DGP		**Link function**	**Density**	**Score**	**Realised information**	**Information**
Type	Distribution		$p(\boldsymbol{y}_t \mid \theta_t)$	$\dfrac{\mathrm{d}\ell(\boldsymbol{y_t} \mid \theta_t)}{\mathrm{d}\theta_t}$	$-\dfrac{\mathrm{d}^2\ell(\boldsymbol{y_t} \mid \theta_t)}{\mathrm{d}\theta_t^2}$	$\mathbb{E}\left[-\dfrac{\mathrm{d}^2\ell(\boldsymbol{y_t} \mid \theta_t)}{\mathrm{d}\theta_t^2}\middle\vert\theta_t\right]$
Count	Poisson	$\lambda_t = \exp(\theta_t)$	$\lambda_t^{y_t}\,\exp(-\lambda_t)/y_t!$	$y_t - \lambda_t$	λ_t	λ_t
Count	Negative bin.	$\lambda_t = \exp(\theta_t)$	$\dfrac{\Gamma(\kappa+y_t)\left(\frac{\kappa}{\kappa+\lambda_t}\right)^{\kappa}\left(\frac{\lambda_t}{\kappa+\lambda_t}\right)^{y_t}}{\Gamma(\kappa)\Gamma(y_t+1)}$	$y_t - \dfrac{\lambda_t(\kappa+y_t)}{\kappa+\lambda_t}$	$\dfrac{\kappa\lambda_t(\kappa+y_t)}{(\kappa+\lambda_t)^2}$	$\dfrac{\kappa\,\lambda_t}{\kappa+\lambda_t}$
Intensity	Exponential	$\lambda_t = \exp(\theta_t)$	$\lambda_t\,\exp(-\lambda_t y_t)$	$1-\lambda_t\,y_t$	$y_t\lambda_t$	1
Duration	Gamma	$\beta_t = \exp(\theta_t)$	$\dfrac{y_t^{\kappa-1}\exp(-y_t/\beta_t)}{\Gamma(\kappa)\beta_t^{\kappa}}$	$\dfrac{y_t}{\beta_t} - \kappa$	$\dfrac{y_t}{\beta_t}$	κ
Duration	Weibull	$\beta_t = \exp(\theta_t)$	$\dfrac{\kappa\,(y_t/\beta_t)^{\kappa-1}}{\beta_t\exp\{(y_t/\beta_t)^{\kappa}\}}$	$\kappa\left(\dfrac{y_t}{\beta_t}\right)^{\kappa} - \kappa$	$\kappa^2\left(\dfrac{y_t}{\beta_t}\right)^{\kappa}$	κ^2
Volatility	Gaussian	$\sigma_t^2 = \exp(\theta_t)$	$\dfrac{\exp\{-y_t^2/(2\sigma_t^2)\}}{\{2\pi\sigma_t^2\}^{1/2}}$	$\dfrac{y_t^2}{2\sigma_t^2} - \dfrac{1}{2}$	$\dfrac{y_t^2}{2\sigma_t^2}$	$\dfrac{1}{2}$
Volatility	Student's t	$\sigma_t^2 = \exp(\theta_t)$	$\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)\left(1+\frac{y_t^2}{(\nu-2)\sigma_t^2}\right)^{-\frac{\nu+1}{2}}}{\sqrt{(\nu-2)\pi}\,\Gamma\left(\nu/2\right)\sigma_t}$	$\dfrac{\omega_t\,y_t^2}{2\sigma_t^2} - \dfrac{1}{2}$ $\omega_t := \dfrac{\nu+1}{\nu-2+y_t^2/\sigma_t^2}$	$\dfrac{\nu-2}{\nu+1}\,\dfrac{\omega_t^2\,y_t^2}{2\sigma_t^2}$	$\dfrac{\nu}{2\nu+6}$
Dependence	Gaussian	$\rho_t = \dfrac{1-\exp(-\theta_t)}{1+\exp(-\theta_t)}$	$\dfrac{\exp\left\{-\frac{y_{1t}^2+y_{2t}^2-2\rho_t y_{1t}y_{2t}}{2(1-\rho_t^2)}\right\}}{2\pi\sqrt{1-\rho_t^2}}$	$\dfrac{\rho_t}{2} + \dfrac{1}{2}\dfrac{z_{1t}\,z_{2t}}{1-\rho_t^2}$ $z_{1t} := y_{1t} - \rho_t y_{2t}$ $z_{2t} := y_{2t} - \rho_t y_{1t}$	$0 \not\leq \dfrac{1}{4}\dfrac{z_{1t}^2+z_{2t}^2}{1-\rho_t^2} - \dfrac{1-\rho_t^2}{4}$	$\dfrac{1+\rho_t^2}{4}$
Dependence	Student's t	$\rho_t = \dfrac{1-\exp(-\theta_t)}{1+\exp(-\theta_t)}$	$\dfrac{\nu\left(1+\frac{y_{1t}^2+y_{2t}^2-2\rho_t y_{1t}y_{2t}}{(\nu-2)(1-\rho_t^2)}\right)^{-\frac{\nu+2}{2}}}{2\pi(\nu-2)\sqrt{1-\rho_t^2}}$	$\dfrac{\rho_t}{2} + \dfrac{\omega_t}{2}\dfrac{z_{1t}\,z_{2t}}{1-\rho_t^2}$ $z_{1t} := y_{1t} - \rho_t y_{2t}$ $z_{2t} := y_{2t} - \rho_t y_{1t}$	$0 \not\leq \dfrac{\omega_t}{4}\dfrac{z_{1t}^2+z_{2t}^2}{1-\rho_t^2} - \dfrac{1-\rho_t^2}{4} - \dfrac{1}{2}\dfrac{\omega_t^2}{\nu+2}\dfrac{z_{1t}^2\,z_{2t}^2}{(1-\rho_t^2)^2}$ $\omega_t := \dfrac{\nu+2}{\nu-2+\frac{y_{1t}^2+y_{2t}^2-2\rho_t y_{1t}y_{2t}}{1-\rho_t^2}}$	$\dfrac{2+\nu(1+\rho_t^2)}{4(\nu+4)}$

Note: The table contains nine data-generating processes (DGPs) and link functions from Koopman et al. (2016). For each model, the DGP is given by the linear Gaussian state equation in combination with the observation density and link functions indicated in the table. The table further displays scores, realised information quantities and expected information quantities. The realised information quantities are nonnegative except for the bottom two models.

True process: We consider Poisson-generated count data $y_t \in \mathbb{N}$ with a time-varying intensity $\lambda_t^\dagger = \exp(\theta_t^\dagger)$, where $\theta_t^\dagger \in \mathbb{R}$ is the time-varying parameter of interest, while the exponential link function ensures $\lambda_t^\dagger > 0$. We assume correct specification of the observation density, i.e. $p^\dagger(y|\lambda) = p(y|\lambda) = \lambda^y/y! \exp(-\lambda)$ for $y \in \mathbb{N}$ and $\lambda > 0$. The true process follows a linear Gaussian state-transition equation, i.e. $\theta_t^\dagger = \omega^\dagger + \phi^\dagger \theta_{t-1}^\dagger + \eta_t$ for $t = 1, \ldots, T$, with $\theta_0^\dagger = 0$, $T = 5000$, $\omega^\dagger = 0$, $\phi^\dagger = 0.98$ and Gaussian disturbances $\eta_t \sim$ i.i.d. $\mathrm{N}(0, \sigma_\eta^2), \forall t$. We vary the value of $\sigma_\eta \geq 0$ and, for each value of σ_η, perform 10^3 replications. As a point of reference, we implement the standard bootstrap particle filter (Malik and Pitt, 2011) using the correct state transition and true (hyper)parameter values, which should give highly accurate—if generally infeasible—state estimates.

ISD filter: We implement the linear prediction (47) in combination with the implicit update (4). The filter is initialized using the true parameter value, i.e. $\theta_{0|0}^{\text{im}} = \theta_0^\dagger = 0$. The score is $\nabla \ell(y|\theta) = y - \exp(\theta)$ for $y \in \mathbb{N}$ and $\theta \in \mathbb{R}$, while the Hessian $\nabla^2 \ell(y|\theta) = -\exp(\theta) < 0$ is strictly negative, such that $\theta \mapsto \ell(y|\theta)$ is strictly concave. Since the negative Hessian $-\nabla^2 \ell(y|\theta) = \exp(\theta)$ does not depend on y, it equals the Fisher information quantity. We take a static learning rate $H_t^{\text{im}} = \eta^{\text{im}} > 0$, where η^{im} is a parameter to be estimated. The global maximizer (4) can be found using standard Newton-Raphson (NR) iterates (6).

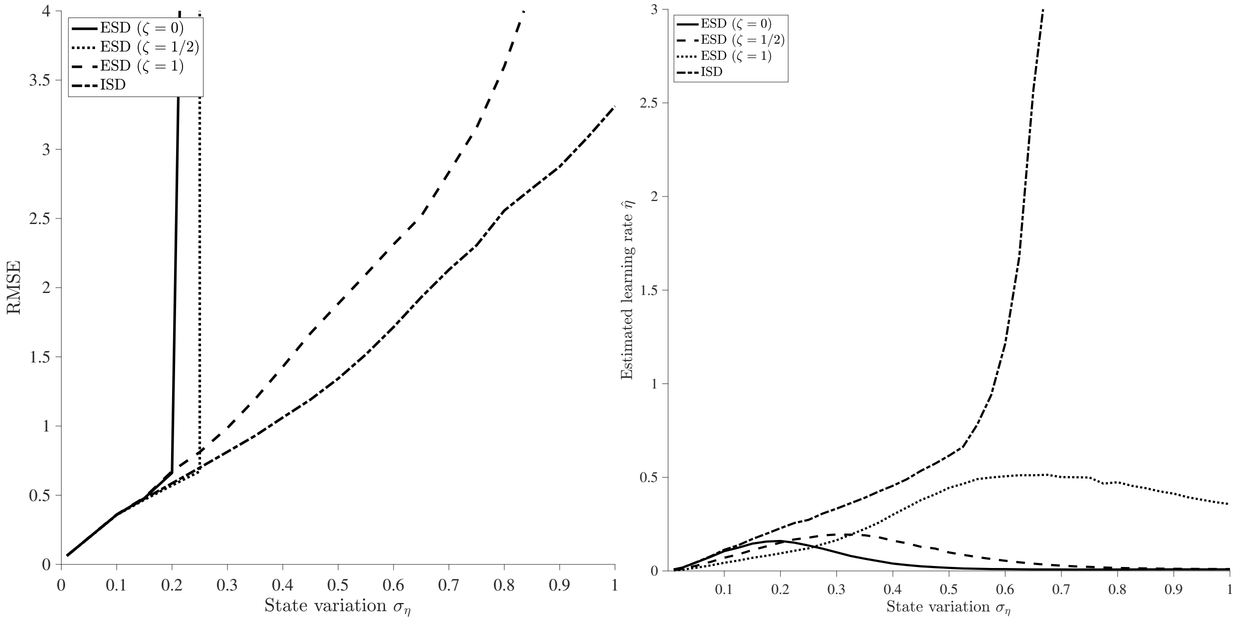
ESD filter: We implement the linear prediction (47) in combination with the explicit update (3). As above, the filter is initialized using the true parameter value, i.e. $\theta_{0|0}^{\text{ex}} = \theta_0^\dagger = 0$. We follow the literature (e.g. Koopman et al., 2016) in taking the learning rate to be $H_t^{\text{ex}} = \eta^{\text{ex}} \exp(-\zeta\, \theta_{t|t-1}^{\text{ex}})$ with $\zeta \in \{0, 1/2, 1\}$, where $\eta^{\text{ex}} > 0$ is a static parameter to be estimated; hence, H_t^{ex} is time-varying unless $\zeta = 0$. The driving force in the explicit filter is $H_t^{\text{ex}} \nabla \ell(y_t|\theta_{t|t-1}^{\text{ex}}) = \eta^{\text{ex}} \exp(-\zeta\, \theta_{t|t-1}^{\text{ex}})(y - \exp(\theta_{t|t-1}^{\text{ex}}))$. This driving force contains exponential terms—except if $y = 0$ and $\zeta = 1$—and hence fails to be Lipschitz in the variable $\theta^{\text{ex}}_{t|t-1}$. Moreover, its average (over y) fails to be Lipschitz irrespective of $\zeta \in \{0, 1/2, 1\}$.

(Hyper)parameter tuning: Both score-driven filters contain three static parameters, which are collected in the (hyper)parameter vector $\boldsymbol{\psi}^j := (\omega^j, \phi^j, \eta^j)'$ for $j \in \{\text{im}, \text{ex}\}$. We extend the standard practice (e.g. Blasques et al., 2023) for explicit filters by computing $\widehat{\boldsymbol{\psi}}^j := \arg\max_{\boldsymbol{\psi}^j} \sum_t \ell(y_t|\theta^j_{t|t-1})$ for $j \in \{\text{im}, \text{ex}\}$, using the in-sample period consisting of the first 2500 data points. As the true process is (believed to be) stationary, we impose $\hat{\phi}^j \in (0, 1)$ for $j \in \{\text{im}, \text{ex}\}$. As the estimated values of ω^j and ϕ^j tend to be very close to $\omega^\dagger$ and $\phi^\dagger$, respectively, we also consider the case where we set $\omega^j = \omega^\dagger$ and $\phi^j = \phi^\dagger$ for $j \in \{\text{im}, \text{ex}\}$. Only a single static (hyper)parameter is then estimated for each filter: the learning-rate parameter $\eta^j > 0$ for $j \in \{\text{im}, \text{ex}\}$. Using the estimated parameters, the whole data set (i.e. including the out-of-sample period) is used to construct filtered paths. We are particularly interested in how the estimate $\hat{\eta}^j$ varies with σ_η.

Results: As Figure 2b illustrates, the estimated learning rate for the ISD filter is monotone increasing in $\sigma_\eta \geq 0$, which controls the variability of the true path $\{\theta^\dagger_t\}$. This is intuitive: as the true-parameter changes are more pronounced, the filter's sensitivity should be increased. The estimated learning rate of the explicit filter, however, achieves a maximum, after which it falls. The reason is that tracking the true path $\{\theta^\dagger_t\}$ becomes increasingly difficult as σ_η is increased, such that large prediction errors occur more frequently; this would lead to the explicit filter to frequently overshoot or diverge unless the learning rate is artificially reduced. This reduction, however, may come at the cost of filtering performance.

Finally, as Figure 2a illustrates, the implicit and explicit filters achieve similar root mean squared errors (RMSEs) in tracking the true state $\{\theta_t^\dagger\}$ for state-variation parameters not exceeding 0.15. Moreover, the explicit filter with $\zeta = 1/2$ demonstrates (slightly) better RMSE performance under modest state variations compared to the other explicit filters, consistent with Koopman et al. (2016). From this point, however, the tracking performance of the explicit filters sharply deteriorates, as can be seen from the RMSE curves either exiting the figure at the top or abruptly vanishing. A closer inspection of the data reveals that this is indeed due to overshooting and/or diverging, depending on the value of $\zeta \in \{0, 1/2, 1\}$. The implicit filter, on the other hand, provides a stable and predictable performance for all values of σ_η.

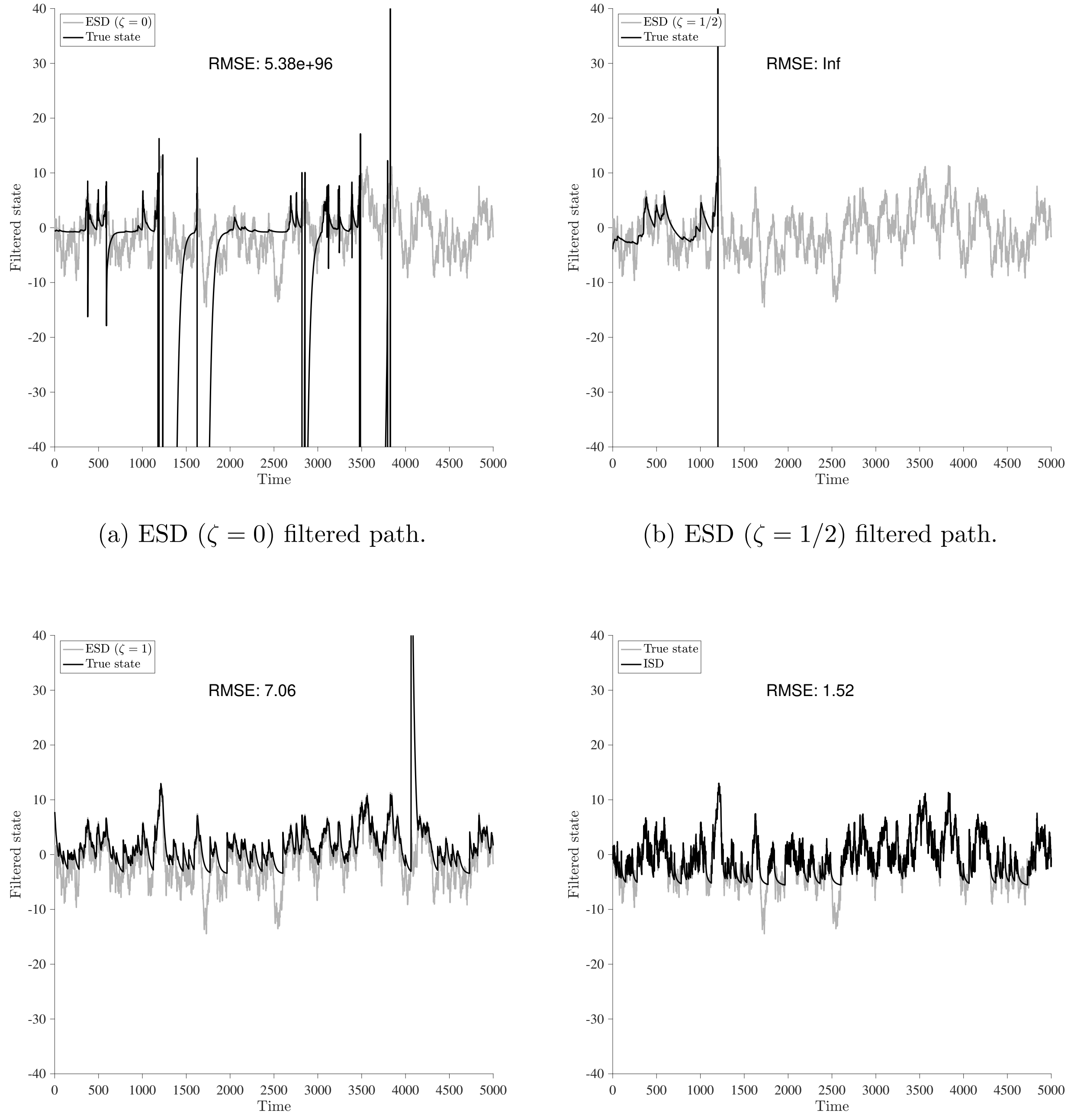
Figure 3 displays a single state path of a Poisson model characterized by state-transition parameters $\sigma_\eta = 0.825$, $\omega^\dagger = -0.01$, and $\phi^\dagger = 0.98$. We filter the true state path using the ESD filter with three choices of scaling for the score and the ISD filter. In Figure 3a, employing identity scaling, the ESD filter frequently exhibits overshooting downward and slow recovery. Notably, around time 3800, it takes over 1000 time points for the filter to approach the true state path. The ESD filter with inverse square root Fisher scaling in Figure 3b performs even worse as it exponentially diverges to infinity in alternating fashion. The ESD filter with inverse Fisher scaling, shown in Figure 3c, is neither capable of reliably filtering the same state path, which is evident from the upward overshooting of the filter. Conversely, the ISD filter shown in Figure 3d performs substantially better, as it neither overshoots nor diverges, but instead quite successfully tracks the true path, with the exception of occasional low state values.



(a) RMSEs of predicted updates $\{\theta^j_{t|t-1}\}$.

(b) Estimated learning rates $\hat{\eta}$

Figure 2: RMSE = root mean squared error. ESD = explicit score-driven. ISD = implicit score-driven. (a) RMSEs of predicted updates $\{\theta^j_{t|t-1}\}$ and (b) estimated learning rates $\hat{\eta}$ while filtering a Poisson model with varying state variations σ_η using the ESD filter with identity scaling ($\zeta = 0$), inverse square root Fisher scaling ($\zeta = 1/2$), inverse Fisher scaling ($\zeta = 1$), and using the ISD filter.



(a) ESD ($\zeta = 0$) filtered path.

(b) ESD ($\zeta = 1/2$) filtered path.

(c) ESD ($\zeta = 1$) filtered path.

(d) ISD filtered path.

Figure 3: ESD = explicit score-driven. ISD = implicit score-driven. Filtering one true state $\{\theta_t^\dagger\}$ path of a Poisson model (grey lines) with state-transition parameters $\sigma_\eta = 0.825$, $\omega^\dagger = -0.01$, and $\phi^\dagger = 0.98$ using the ESD filter with (a) identity scaling ($\zeta = 0$), (b) inverse square root Fisher scaling ($\zeta = 1/2$), (c) inverse Fisher scaling ($\zeta = 1$), and (d) using the ISD filter (black lines). Figures (a), (b), (c), and (d) show the same simulated path.

6 Conclusion

Tracking latent time-varying parameters in the presence of possible model misspecification is challenging, particularly when the true parameters exhibit large fluctuations and/or non-stationary dynamics. We derived performance guarantees for score-driven filters by presenting upper bounds for long-run root mean squared filtering errors. We distinguished between two classes of filters: *explicit* score-driven (ESD) and *implicit* score-driven (ISD). While the first class contains all score-driven filters in the literature, known variously as dynamic conditional score (DCS) or generalized autoregressive score (GAS) filters, the second class is essentially new. We relaxed conditions on the true parameter process considerably compared to recent work on error bounds for tracking latent time-varying parameters. These studies typically impose a limit on the true parameter variation, thereby excluding many realistic data generating processes, such as linear and Gaussian dynamics relevant to, for example, the Kalman filter. In contrast, we only necessitate a finite second moment for the (pseudo-)true parameter increments over time. Our theoretical analysis revealed, for the first time, that ESD filters require regularity conditions on the researcher-postulated logarithmic observation density. Specifically, Lipschitz continuity of the gradient, or equivalently, β-smoothness, is required to prevent the ESD filter from frequently 'overshooting' and, possibly, diverging to infinity. In contrast, ISD filters do not require this restrictive regularity condition. Indeed, our simulation studies across a wide variety of settings demonstrated that, when the true-parameter process is quite volatile, the ISD filter successfully tracks the true parameter even when the ESD filter diverges.

References

Artemova, M., F. Blasques, J. van Brummelen, and S. J. Koopman (2022a). Score-driven models: Methodology and theory. In *Oxford Research Encyclopedia of Economics and Finance*.

Artemova, M., F. Blasques, J. van Brummelen, and S. J. Koopman (2022b). Score-driven models: Methods and applications. In *Oxford Research Encyclopedia of Economics and Finance*.

Ascher, U. M., S. J. Ruuth, and B. T. Wetton (1995). Implicit-explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis 32*(3), 797–823.

Blasques, F., A. Harvey, S. Koopman, and A. Lucas (2023). Time-varying parameters in econometrics: The editor's foreword. *Journal of Econometrics*.

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

Cao, X., J. Zhang, and H. V. Poor (2019). On the time-varying distributions of online stochastic optimization. In *2019 American Control Conference (ACC)*, pp. 1494–1500. IEEE.

Cox, D. R., G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115.

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*(5), 777–795.

Durbin, J. and S. J. Koopman (2012). *Time series analysis by state space methods*, Volume 38. OUP Oxford.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association 87*(418), 501–509.

Gorgi, P., C. Lauria, and A. Luati (2023). On the optimality of score-driven models. *Biometrika*, asad067.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman Filter*. CUP.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, Volume 52. Cambridge University Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*(1), 35–45.

Koopman, S. J., A. Lucas, and M. Scharth (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics 98*(1), 97–110.

Lanconelli, A. and C. S. Lauria (2023). Maximum likelihood with a time varying parameter. *arXiv preprint arXiv:2302.14529*.

Lange, R.-J. (2020). Bellman filtering for state-space models. *arXiv preprint arXiv:2008.11477*.

Lange, R.-J., B. van Os, and D. J. van Dijk (2022). Robust observation-driven models using proximal-parameter updates. *Available at SSRN 4227958*.

Malik, S. and M. K. Pitt (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics 165*(2), 190–209.

Moulines, E. and F. Bach (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 24. Curran Associates, Inc.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Ryu, E. K. and S. Boyd (2014). Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*.

Simonetto, A., E. Dall'Anese, S. Paternain, G. Leus, and G. B. Giannakis (2020). Time-varying convex optimization: Time-structured algorithms and applications. *Proceedings of the IEEE 108*(11), 2032–2048.

Toulis, P., E. Airoldi, and J. Rennie (2014). Statistical analysis of stochastic gradient methods for generalized linear models. In *International Conference on Machine Learning*, pp. 667–675.

Toulis, P. and E. M. Airoldi (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics 45*, 1694–1727.

Toulis, P., T. Horel, and E. M. Airoldi (2021). The proximal Robbins–Monro method. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(1), 188–212.

Wilson, C., V. V. Veeravalli, and A. Nedić (2018). Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control 64*(2), 496–509.

Appendix

A Proof Theorem 1

Here, we derive an upper bound on the long-run RMSE, and corresponding contraction condition for the *ISD* filter under potential misspecification. Suppose Assumptions 1a, 2a, and 3 hold. Start with the ISD filter's update step, which uses differentiability of the log-observation density:

$$\boldsymbol{\theta}_{t|t} = \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}_t^{-1} \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) \tag{21}$$

Take $\boldsymbol{P}_t^{-1} \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t})$ to the left-hand side, pre-multiply both sides by $\boldsymbol{P}_t^{\frac{1}{2}}$, subtract from both sides $\boldsymbol{P}_t^{\frac{1}{2}} \boldsymbol{\theta}_t^\star - \boldsymbol{P}_t^{-\frac{1}{2}} \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)$, and rewrite, to obtain:

$$\boldsymbol{P}_t^{\frac{1}{2}} (\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star) - \boldsymbol{P}_t^{-\frac{1}{2}} (\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)) = \boldsymbol{P}_t^{\frac{1}{2}} (\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star) + \boldsymbol{P}_t^{-\frac{1}{2}} \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star). \tag{22}$$

Compute the quadratic norm on both sides, to obtain:

$$\begin{aligned}
&\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2 + \|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}_t^{-1}}^2 - 2\langle \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star), \boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star \rangle = \\
&\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2 + \|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}_t^{-1}}^2 + 2\langle \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star) \rangle
\end{aligned}$$

Using that by α-strong concavity $\langle \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star), \boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t|t} \rangle \geq \alpha \left\| \boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star \right\|^2$ and subsequently that $\|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}_t^{-1}}^2 \geq \frac{\alpha^2}{\mu_{\max}} \left\| \boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star \right\|^2$:

$$\begin{aligned}
&\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2 + \|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}_t^{-1}}^2 + 2\alpha \|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2 \leq \\
&\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2 + \|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}_t^{-1}}^2 + 2\langle \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star) \rangle \Rightarrow \\
&\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2 + \frac{\alpha^2}{\mu_{\max}} \|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2 + 2\alpha \|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2 \leq \\
&\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2 + \|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}_t^{-1}}^2 + 2\langle \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star) \rangle
\end{aligned}$$

Let $\mathrm{E}_{1:t}[\cdot] := \mathrm{E}_{\boldsymbol{y}_1\ldots\boldsymbol{y}_t;\boldsymbol{\theta}_1^\dagger\ldots\boldsymbol{\theta}_t^\dagger}[\cdot] = \int_{\boldsymbol{\theta}_t^\star}\int_{\boldsymbol{y}_t}\int_{\boldsymbol{\theta}_{t-1}^\star}\int_{\boldsymbol{y}_{t-1}}\cdots\left[\int_{\boldsymbol{\theta}_1^\star}\left[\int_{\boldsymbol{y}_1}(\cdot)p^\dagger(\boldsymbol{y}\mid\boldsymbol{\theta}_t^\dagger)d\boldsymbol{y}\right]p_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta}\right]d\boldsymbol{y}d\boldsymbol{\theta}\ldots.d\boldsymbol{y}d\boldsymbol{\theta}$ denote the expectation with respect to the true state and observation path until time t, where first the expectation is taken w.r.t. $\boldsymbol{\theta}_t^\dagger$, then $\boldsymbol{y}_t$, $\boldsymbol{\theta}_{t-1}^\dagger$, $\boldsymbol{y}_{t-1}$ etc. Note that the pseudo-true parameter is some unknown function $f(\cdot)$ of the true parameter: $\boldsymbol{\theta}_t^\star = f(\boldsymbol{\theta}_t^\dagger)$. Now we take the expectation with respect to the true state and observation path until time t on both sides and combine terms:

$$\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}_t+(2\alpha+\frac{\alpha^2}{\mu_{\max}})\boldsymbol{I}_d}] \leq \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}_t}] + \mathrm{E}_{1:t}[\|\nabla(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|^2_{\boldsymbol{P}_t^{-1}}] \tag{23}$$

Here, we used that $\mathrm{E}_{1:t}[\langle\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\rangle] = 0$, as $\mathrm{E}_{1:t}[\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)] = 0$. Now, using Assumption 3 and that $\boldsymbol{P}_t$ is assumed to be positive definite:

$$\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}_t+(2\alpha+\frac{\alpha^2}{\mu_{\max}})\boldsymbol{I}_d}] \leq \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}_t}] + \frac{\sigma^2_{\max}}{\mu_{\min}} \tag{24}$$

Hence, the MSE is contractive unless the prediction is within the noise-dominated region (NDR). In the next step, we get slackness of the bound.

$$\underbrace{\lambda_{\min}(\boldsymbol{P}_t + (2\alpha + \frac{\alpha^2}{\mu_{\max}})\boldsymbol{I}_d)}_{= \mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}} \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2] \leq \mu_{\max}\,\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2] + \frac{\sigma^2_{\max}}{\mu_{\min}}$$

$$\underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2]}_{=:\ \mathrm{MSE}_{t|t}} \leq \underbrace{\frac{\mu_{\max}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}}}_{>0\text{, we can choose } \boldsymbol{P}_t \text{ s.t. this term } <1} \underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2]}_{=:\ \mathrm{MSE}_{t|t-1}} + \underbrace{\frac{\frac{\sigma^2_{\max}}{\mu_{\min}}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}}}_{>0} \tag{25}$$

As we are interested in *root* mean squared filtering errors over time, we take the square root on both sides, and use that $\sqrt{z_1 + z_2} \leq \sqrt{z_1} + \sqrt{z_2}$, for non-negative scalars z_1 and z_2, to obtain:

$$\mathrm{RMSE}_{t|t} \leq \sqrt{\frac{\mu_{\max}}{\mu_{\min}+2\alpha+\frac{\alpha^2}{\mu_{\max}}}}\mathrm{RMSE}_{t|t-1} + \sqrt{\frac{\frac{\sigma^2_{\max}}{\mu_{\min}}}{\mu_{\min}+2\alpha+\frac{\alpha^2}{\mu_{\max}}}} \tag{26}$$

$$\begin{aligned}
\mathrm{RMSE}_{t|t-1} &= \sqrt{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}^\star_t\|^2]} \\
&= \sqrt{\mathrm{E}_{1:t}[\|\underbrace{\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}^\star_{t-1}}_{=:\,\boldsymbol{z}_1}\underbrace{-(\boldsymbol{\theta}^\star_t-\boldsymbol{\theta}^\star_{t-1})}_{=:\,\boldsymbol{z}_2}\|^2]} \\
&\leq \sqrt{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}^\star_{t-1}\|^2]} + \underbrace{\sqrt{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}^\star_t-\boldsymbol{\theta}^\star_{t-1}\|^2]}}_{=:\,q} \\
&= \sqrt{\mathrm{E}_{1:t-1}[\|\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}^\star_{t-1}\|^2]} + q \\
&= \sqrt{\mathrm{E}_{1:t-1}[\|\underbrace{\boldsymbol{\Phi}(\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}^\star_{t-1})}_{=:\,\boldsymbol{z}_1}+\underbrace{\boldsymbol{\omega}+(\boldsymbol{\Phi}-\boldsymbol{I}_d)\boldsymbol{\theta}^\star_{t-1}}_{=:\,\boldsymbol{z}_2}\|^2]} + q \\
&\leq \sqrt{\mathrm{E}_{1:t-1}[\|\boldsymbol{\Phi}(\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}^\star_{t-1})\|^2]} + q + \sqrt{\mathrm{E}_{1:t-1}[\|\underbrace{\boldsymbol{\omega}}_{=:\,\boldsymbol{z}_1}+\underbrace{(\boldsymbol{\Phi}-\boldsymbol{I}_d)\boldsymbol{\theta}^\star_{t-1}}_{=:\,\boldsymbol{z}_2}\|^2]} \\
&\leq \sqrt{E_{1:t-1}\left[\|\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}^\star_{t-1}\|^2_{\boldsymbol{\Phi}'\boldsymbol{\Phi}}\right]} + q + \sqrt{\mathrm{E}_{1:t-1}\left[\|\boldsymbol{\omega}\|^2\right]} + \sqrt{\mathrm{E}_{1:t-1}\left[\|(\boldsymbol{\Phi}-\boldsymbol{I}_d)\boldsymbol{\theta}^\star_{t-1}\|^2\right]} \\
&\leq \sqrt{\lambda_{\max}(\boldsymbol{\Phi}'\boldsymbol{\Phi})\,\mathrm{E}_{1:t-1}\left[\|\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}^\star_{t-1}\|^2\right]} + q + \sqrt{\boldsymbol{\omega}'\boldsymbol{\omega}} + \sqrt{\mathrm{E}_{1:t-1}\left[\|\boldsymbol{\theta}^\star_{t-1}\|^2_{(\boldsymbol{\Phi}-\boldsymbol{I}_d)'(\boldsymbol{\Phi}-\boldsymbol{I}_d)}\right]} \\
&= \|\boldsymbol{\Phi}\|_2\underbrace{\sqrt{\mathrm{E}_{1:t-1}\left[\|\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}^\star_{t-1}\|^2\right]}}_{=:\,\mathrm{RMSE}_{t-1|t-1}} + q + \|\boldsymbol{\omega}\| + \|(\boldsymbol{\Phi}-\boldsymbol{I}_d)\|_2\underbrace{\sqrt{\mathrm{E}_{1:t-1}\left[\|\boldsymbol{\theta}^\star_{t-1}\|^2\right]}}_{\leq s}
\end{aligned}$$

where we used Hölders inequality for random vectors with $p = q = 2$ in the first, second, and third inequality. That is, for the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for $1 < p, q < \infty$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, and for real- or complex-valued random vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ on Ω, with i-th components z_1^i and z_2^i, respectively, Hölder's inequality states:

$$\mathbb{E}\left[\sum_{i=1}^n \int_\Omega \left|z_1^i(\omega)z_2^i(\omega)\right| \mathrm{d}\mathbb{P}\right] \leq \mathbb{E}\left[\sum_{i=1}^n \int_\Omega \left|z_1^i(\omega)z_2^i(\omega)\right|^p \, \mathrm{d}\mathbb{P}\right]^{1/p} \mathbb{E}\left[\sum_{i=1}^n \int_\Omega \left|z_1^i(\omega)z_2^i(\omega)\right|^q \, \mathrm{d}\mathbb{P}\right]^{1/q}$$

Hence:

$\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_2] \leq \mathrm{E}[|\boldsymbol{z}_1'\boldsymbol{z}_2|] \leq \sqrt{\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_1][\boldsymbol{z}_2'\boldsymbol{z}_2]}$ By Hölders inequality for random vectors.

$$\sqrt{\mathrm{E}[(\boldsymbol{z}_1+\boldsymbol{z}_2)'(\boldsymbol{z}_1+\boldsymbol{z}_2)]} = \sqrt{\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_1] + \mathrm{E}[\boldsymbol{z}_2'\boldsymbol{z}_2] + 2\,\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_2]}$$

$$\leq \sqrt{\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_1] + \mathrm{E}[\boldsymbol{z}_2'\boldsymbol{z}_2] + 2\sqrt{\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_1][\boldsymbol{z}_2'\boldsymbol{z}_2]}} = \sqrt{(\sqrt{\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_1]} + \sqrt{\mathrm{E}[\boldsymbol{z}_2'\boldsymbol{z}_2]})^2}$$

$$= \sqrt{\mathrm{E}[\boldsymbol{z}_1'\boldsymbol{z}_1]} + \sqrt{\mathrm{E}[\boldsymbol{z}_2'\boldsymbol{z}_2]}$$

Combining:

$$\mathrm{RMSE}_{t|t} \leq \underbrace{\sqrt{\frac{\mu_{\max}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}}}}_{=:\ c}\mathrm{RMSE}_{t|t-1} + \underbrace{\sqrt{\frac{\frac{\sigma^2_{\max}}{\mu_{\min}}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}}}}_{=:\ d} \tag{27}$$

$$\mathrm{RMSE}_{t|t-1} \leq \underbrace{\|\boldsymbol{\Phi}\|_2}_{=:\ a}\mathrm{RMSE}_{t-1|t-1} + \underbrace{q + \|\boldsymbol{\omega}\| + \|(\boldsymbol{I}_d - \boldsymbol{\Phi})\|_2\, s}_{=:\ b} \tag{28}$$

Repeated substitution of the recursions (32) and (33) yields:

$$\begin{aligned}\mathrm{RMSE}_{t|t} &\leq c^t a^{t-1}\mathrm{RMSE}_{1|0} + d\sum_{i=0}^{t-1}(ca)^i + bc\sum_{i=0}^{t-2}(ca)^i \\ &= c^t a^{t-1}\mathrm{RMSE}_{1|0} + d\frac{1-(ca)^t}{1-ca} + bc\frac{1-(ca)^{t-1}}{1-ca}, \quad ca \neq 1\end{aligned}$$

where use a geometric series result: $\sum_{i=0}^{t} x^i = \left(1 - x^{t+1}\right)/(1-x)$ for $x \neq 1$. Moreover, under the following condition, which we refer to as the *contraction condition*:

$$\sqrt{\frac{\mu_{\max}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}}}\|\boldsymbol{\Phi}\|_2 < 1 \tag{29}$$

The sums converge as $t \to \infty$, i.e. we can upper bound the *long-run root mean squared filtering error*:

$$\limsup_{t\to\infty}\mathrm{RMSE}_{t|t} \leq \frac{\frac{\sigma_{\max}}{\sqrt{\mu_{\min}}} + \sqrt{\mu_{\max}}\left[q + \|\boldsymbol{\omega}\| + \|\boldsymbol{\Phi} - \boldsymbol{I}_d\|_2\, s\right]}{\sqrt{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}} - \sqrt{\mu_{\max}}\|\boldsymbol{\Phi}\|_2}. \tag{30}$$

Next, we derive an upper bound on the long-run RMSE, and corresponding contraction condition for the *ESD* filter under potential misspecification. Suppose Assumptions 1a, 2b and 3 hold. Start with $\boldsymbol{\theta}_{t|t} = \boldsymbol{\theta}_{t|t}$, then subtract the true state $\boldsymbol{\theta}_t^\star$ on both sides, subtract $\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_{t|t-1}$ from the right side, and pre-multiply both sides by $\boldsymbol{P}_t^{\frac{1}{2}}$:

$$\boldsymbol{P}_t^{\frac{1}{2}}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star) = \boldsymbol{P}_t^{\frac{1}{2}}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_{t|t-1} + \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star)$$

Compute the quadratic norm on both sides, to obtain:

$$\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^*\|_{\boldsymbol{P}}^2 = \|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_{t|t-1}\|_{\boldsymbol{P}}^2 + \|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^*\|_{\boldsymbol{P}}^2 + 2\langle \boldsymbol{P}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_{t|t-1}), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^*\rangle$$

Substituting $\boldsymbol{P}_t(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_{t|t-1}) = \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})$ and $\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_{t|t-1} = \boldsymbol{P}_t^{-1}\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})$ from the ESD filter's prediction step, and taking the with respect to the true state and observation path until time t, we get:

$$\begin{aligned}
&\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2] \\
&= \mathrm{E}_{1:t}[\|\boldsymbol{P}_t^{-1}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|_{\boldsymbol{P}_t}^2] + \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2] + 2\,\mathrm{E}_{1:t}[\langle \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\rangle] \\
&= \mathrm{E}_{1:t}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|_{\boldsymbol{P}_t^{-1}}^2] + \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2] - 2\,\mathrm{E}_{1:t}[\langle \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star), \boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t|t-1}\rangle] \\
&\leq \mathrm{E}_{1:t}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|_{\boldsymbol{P}_t^{-1}}^2] + \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{P}_t}^2] - 2\alpha\,\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2]
\end{aligned}$$

Here, we used that $\mathrm{E}_{1:t}[\langle \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\rangle] = 0$, as $\mathrm{E}_{1:t}[\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)] = 0$, by subtracting it from the right side. In the last step, we used that the log-likelihood $\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta})$ is once continuously differentiable and strongly concave in $\boldsymbol{\theta}$ with parameter α (Assumption 2a). After adding weight matrices and bounding the terms using the eigenvalues of the weight matrices, we get:

$$\mu_{\min} \underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^{\star}\|^2]}_{=:\ \mathrm{MSE}_{t|t}} \leq \mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^{\star}\|^2_{\boldsymbol{P}_t}]$$

$$\leq \underbrace{\lambda_{\max}(\boldsymbol{P}_t - 2\alpha \boldsymbol{I}_d)}_{=\,\mu_{\max} - 2\alpha} \underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^{\star}\|^2]}_{=:\ \mathrm{MSE}_{t|t-1}} + \underbrace{\lambda_{\max}(\boldsymbol{P}_t^{-1})}_{=\,\frac{1}{\mu_{\min}}} \mathrm{E}_{1:t}[\|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|^2]$$

Rewrite:

$$\mathrm{MSE}_{t|t} \leq \frac{\mu_{\max} - 2\alpha}{\mu_{\min}} \mathrm{MSE}_{t|t-1} + \frac{\mathrm{E}_{1:t}[\|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|^2]}{\mu_{\min}^2} \tag{31}$$

As we are interested in the *root* mean squared filtering errors over time, we twice use $\sqrt{z_1 + z_2} \leq \sqrt{z_1} + \sqrt{z_2}$, for non-negative scalar z_1 and z_2, subsequently Hölders inequality for random vectors, and in the last inequality we use that the log-likelihood $\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta})$ is once continuously differentiable and β-smooth in $\boldsymbol{\theta}$ with parameter β (Assumption 2b) and bounded information (Assumption 3), to obtain:

$$\begin{aligned}
&\mathrm{RMSE}_{t|t} \\
&\leq \sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} \mathrm{RMSE}_{t|t-1} + \frac{1}{\mu_{\min}} \sqrt{\mathrm{E}_{1:t}[\|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|^2]} \\
&\leq \sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} \mathrm{RMSE}_{t|t-1} + \frac{1}{\mu_{\min}} \left(\sqrt{E_{1:t}[\|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^{\star})\|^2]} + \sqrt{E_{1:t}[\|\nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^{\star})\|^2]} \right) \\
&\leq \sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} =:\ \mathrm{RMSE}_{t|t-1} + \frac{1}{\mu_{\min}} \left(\beta \underbrace{\sqrt{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^{\star}\|^2]}}_{\mathrm{RMSE}_{t|t-1}} + \sigma_{\max} \right)
\end{aligned}$$

We note that the "(one-step) predicted RMSE" of the ESD filter is equivalent to that of the ISD filter, thus after combining, we get:

$$\mathrm{RMSE}_{t|t} \leq \underbrace{\left(\sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} + \frac{\beta}{\mu_{\min}}\right)}_{=:\ c} \mathrm{RMSE}_{t|t-1} + \underbrace{\frac{\sigma_{\max}}{\mu_{\min}}}_{=:\ d} \tag{32}$$

$$\mathrm{RMSE}_{t|t-1} \leq \underbrace{\|\boldsymbol{\Phi}\|_2}_{=:\ a} \mathrm{RMSE}_{t-1|t-1} + \underbrace{q + \|\boldsymbol{\omega}\| + \|(\boldsymbol{I}_d - \boldsymbol{\Phi})\|_2\, s}_{=:\ b} \tag{33}$$

Repeated substitution of the recursions (32) and (33) yields:

$$\begin{aligned}\mathrm{RMSE}_{t|t} &\leq c^t a^{t-1} \mathrm{RMSE}_{1|0} + d \sum_{i=0}^{t-1} (ca)^i + bc \sum_{i=0}^{t-2} (ca)^i \\ &= c^t a^{t-1} \mathrm{RMSE}_{1|0} + d \frac{1 - (ca)^t}{1 - ca} + bc \frac{1 - (ca)^{t-1}}{1 - ca}, \quad ca \neq 1\end{aligned}$$

where use a geometric series result: $\sum_{i=0}^{t} x^i = \left(1 - x^{t+1}\right)/(1 - x)$ for $x \neq 1$. Moreover, under the following condition, which we refer to as the *contraction condition*:

$$\left(\sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} + \frac{\beta}{\mu_{\min}}\right) \|\boldsymbol{\Phi}\|_2 < 1 \tag{34}$$

The sums converge as $t \to \infty$, i.e. we can upper bound the *long-run root mean squared filtering error*:

$$\limsup_{t \to \infty} \mathrm{RMSE}_{t|t} \leq \frac{\frac{\sigma_{\max}}{\sqrt{\mu_{\min}}} + \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}}\right) \left[q + \|\boldsymbol{\omega}\| + \|\boldsymbol{\Phi} - \boldsymbol{I}_d\|_2\, s\right]}{\sqrt{\mu_{\min}} - \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}}\right) \|\boldsymbol{\Phi}\|_2}. \tag{35}$$

B Proof Corollary 1.1

Here, we present an upper bound on the long-run RMSE, and corresponding contraction condition for the *ISD* filter under potential misspecification, when the penalty matrix is a scalar multiple of the identity matrix. Suppose Assumptions 1a, 2a, and 3 hold. As $\boldsymbol{P}_t = \gamma \boldsymbol{I}_d$, this implies that $\mu_{\min} = \mu_{\max} = \gamma$. Then the contraction condition (29) can be simplified as follows: $\sqrt{\frac{\gamma}{\gamma+2\alpha+\frac{\alpha^2}{\gamma}}}\|\mathbf{\Phi}\|_2 < 1 \Leftrightarrow \sqrt{\frac{\gamma^2}{\gamma^2+2\alpha\gamma+\alpha^2}}\|\mathbf{\Phi}\|_2 < 1 \Leftrightarrow \sqrt{\frac{\gamma^2}{(\gamma+\alpha)^2}}\|\mathbf{\Phi}\|_2 < 1 \Leftrightarrow \frac{\gamma}{\gamma+\alpha}\|\mathbf{\Phi}\|_2 < 1$. As we have imposed the penalty to be positive ($\gamma > 0$), the observation log-density is assumed to be strongly concave ($\alpha > 0$), the contraction condition is satisfied for any non-explosive autoregressive matrix ($\|\mathbf{\Phi}\|_2 \leq 1$). For this choice of penalty matrix, the upper bound on the long-run RMSE of the ISD filter (30) is minimal, and simplifies to:

$$\limsup_{t\to\infty} \mathrm{RMSE}_{t|t} \leq \frac{\sigma_{\max} + \gamma\,[q + \|\boldsymbol{\omega}\| + \|\mathbf{\Phi} - \boldsymbol{I}_d\|_2\, s]}{\gamma(1 - \|\mathbf{\Phi}\|_2) + \alpha}. \tag{36}$$

For the RMSE bound and contraction condition of the *ESD* filter when $\boldsymbol{P}_t = \gamma \boldsymbol{I}_d$, we additionally assume $\beta-$smoothness of the observation log-density, i.e. Assumption 2b holds. The contraction condition (34) simplifies to: $(\sqrt{\frac{\gamma-2\alpha}{\gamma}} + \frac{\beta}{\gamma})\|\mathbf{\Phi}\|_2 < 1$. For this choice of penalty matrix, the upper bound on the long-run RMSE of the ESD filter (35) is minimal, and simplifies to:

$$\limsup_{t\to\infty} \mathrm{RMSE}_{t|t} \leq \frac{\sigma_{\max} + \left(\sqrt{\gamma^2 - 2\alpha\gamma} + \beta\right)[q + \|\boldsymbol{\omega}\| + \|\mathbf{\Phi} - \boldsymbol{I}_d\|_2\, s]}{\gamma - \left(\sqrt{\gamma^2 - 2\alpha\gamma} + \beta\right)\|\mathbf{\Phi}\|_2}. \tag{37}$$

C Proof Theorem 2

Here, we derive an upper bound on the long-run RMSE, and corresponding contraction condition for the *ISD* filter. Suppose now that next to Assumptions 2a and 3, now 1b holds (instead of 1a), i.e. the model is correctly specified, and the state-transition equation is linear and Gaussian with known coefficients:

$$\boldsymbol{\theta}_t^\dagger = \boldsymbol{\omega}^\dagger + \mathbf{\Phi}^\dagger \boldsymbol{\theta}_{t-1}^\dagger + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \text{i.i.d. N}(\mathbf{0}, \boldsymbol{Q}^\dagger), \quad \rho(\mathbf{\Phi}^\dagger) \leq 1, \quad \sigma_\eta^2 := \mathrm{tr}(\boldsymbol{Q}^\dagger) < \infty. \tag{38}$$

We use the ISD filter, evaluated in the true = pseudo-true parameters, that is:

$$\text{prediction step:} \qquad \boldsymbol{\theta}_{t|t-1}^{\text{im}} = \boldsymbol{\omega}^\star + \mathbf{\Phi}^\star \boldsymbol{\theta}_{t-1|t-1}^{\text{im}}, \tag{39}$$

$$\text{implicit-gradient update:} \qquad \boldsymbol{\theta}_{t|t}^{\text{im}} = \boldsymbol{\theta}_{t|t-1}^{\text{im}} + \boldsymbol{H}_t \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}^{\text{im}}). \tag{40}$$

The "(one-step) filtered MSE bound" is equivalent to that using Assumption 1a, i.e.:

$$\underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}_t^\star\|^2]}_{=:\ \mathrm{MSE}_{t|t}} \leq \underbrace{\frac{\mu_{\max}}{\mu_{\min}+2\alpha+\frac{\alpha^2}{\mu_{\max}}}}_{>0,\text{ we can choose }\boldsymbol{P}_t\text{ s.t. this term }<1} \underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}_t^\star\|^2]}_{=:\ \mathrm{MSE}_{t|t-1}} + \underbrace{\frac{\frac{\sigma^2_{\max}}{\mu_{\min}}}{\mu_{\min}+2\alpha+\frac{\alpha^2}{\mu_{\max}}}}_{>0}. \tag{41}$$

The "(one-step) predicted MSE bound", is however different. To see this, we substitute the prediction step (47) and implicit update step (48) in $\mathrm{MSE}_{t|t-1}$:

$$\begin{aligned}\underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}_t^\star\|^2]}_{=:\ \mathrm{MSE}_{t|t-1}} &= \mathrm{E}_{1:t}[\|\boldsymbol{\Phi}^\star(\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}_{t-1}^\star)-\boldsymbol{\eta}_t\|^2] \\ &= \mathrm{E}_{1:t-1}[\|\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}_{t-1}^\star\|^2_{\boldsymbol{\Phi}^{\star\prime}\boldsymbol{\Phi}^\star}] + \underbrace{\mathrm{E}_{1:t}[\|\boldsymbol{\eta}_t\|^2]}_{=\mathrm{tr}(\boldsymbol{Q}^\dagger)=:\ \sigma^2_\eta} -2\,\mathrm{E}_{1:t}[\langle\boldsymbol{\Phi}^\star(\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}_{t-1}^\star),\boldsymbol{\eta}_t\rangle] \\ &\leq \underbrace{\lambda_{\max}(\boldsymbol{\Phi}^{\star\prime}\boldsymbol{\Phi}^\star)}_{=\|\boldsymbol{\Phi}^\star\|_2^2} \underbrace{\mathrm{E}_{1:t-1}[\|\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}_{t-1}^\star\|^2]}_{=:\mathrm{MSE}_{t-1|t-1}} + \sigma^2_\eta.\end{aligned}$$

where we have used that $\mathrm{E}_{1:t}[\langle\boldsymbol{\Phi}^\star(\boldsymbol{\theta}_{t-1|t-1}-\boldsymbol{\theta}_{t-1}^\star),\boldsymbol{\eta}_t\rangle]=0$. Combining yields:

$$\mathrm{MSE}_{t|t} \leq \underbrace{\frac{\mu_{\max}}{\mu_{\min}+2\alpha+\frac{\alpha^2}{\mu_{\max}}}}_{=:\ c} \mathrm{MSE}_{t|t-1} + \underbrace{\frac{\frac{\sigma^2_{\max}}{\mu_{\min}}}{\mu_{\min}+2\alpha+\frac{\alpha^2}{\mu_{\max}}}}_{=:\ d} \tag{42}$$

$$\mathrm{MSE}_{t|t-1} \leq \underbrace{\|\boldsymbol{\Phi}^\star\|_2^2}_{=:\ a} \mathrm{MSE}_{t-1|t-1} + \underbrace{\sigma^2_\eta}_{=:\ b} \tag{43}$$

Repeated substitution of the recursions (42) and (43) yields:

$$\begin{aligned}\mathrm{MSE}_{t|t} &\leq c^t a^{t-1}\mathrm{MSE}_{1|0} + d\sum_{i=0}^{t-1}(ca)^i + bc\sum_{i=0}^{t-2}(ca)^i \\ &= c^t a^{t-1}\mathrm{MSE}_{1|0} + d\frac{1-(ca)^t}{1-ca} + bc\frac{1-(ca)^{t-1}}{1-ca}, \quad ca\neq 1\end{aligned}$$

where use a geometric series result: $\sum_{i=0}^t x^i = (1-x^{t+1})/(1-x)$ for $x\neq 1$. Moreover, under the following condition, which we refer to as the *contraction condition*:

$$\frac{\mu_{\max}}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}}} \|\mathbf{\Phi}^\star\|_2^2 < 1, \tag{44}$$

the sums converge as $t \to \infty$, i.e. we upper bound the *long-run mean squared filtering error*:

$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\frac{\sigma_{\max}^2}{\mu_{\min}} + \mu_{\max}\sigma_\eta^2}{\mu_{\min} + 2\alpha + \frac{\alpha^2}{\mu_{\max}} - \mu_{\max}\|\mathbf{\Phi}^\star\|_2}. \tag{45}$$

Next, we derive an upper bound on the long-run RMSE, and corresponding contraction condition for the *ESD* filter. Suppose in addition to Assumptions 2b and 3, now Assumption 1b holds (instead of 1a), i.e. the model is correctly specified, and the state-transition equation is linear and Gaussian with known coefficients:

$$\boldsymbol{\theta}_t^\dagger = \boldsymbol{\omega}^\dagger + \mathbf{\Phi}^\dagger \boldsymbol{\theta}_{t-1}^\dagger + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \text{i.i.d. N}(\mathbf{0}, \boldsymbol{Q}^\dagger), \quad \rho(\mathbf{\Phi}^\dagger) \leq 1, \quad \sigma_\eta^2 := \mathrm{tr}(\boldsymbol{Q}^\dagger) < \infty. \tag{46}$$

We use the ESD filter, evaluated in the true = pseudo-true parameters, that is:

$$\text{prediction step:} \qquad \boldsymbol{\theta}_{t|t-1}^{\text{ex}} \;=\; \boldsymbol{\omega}^\star \;+\; \mathbf{\Phi}^\star\, \boldsymbol{\theta}_{t-1|t-1}^{\text{ex}}, \tag{47}$$

$$\text{implicit-gradient update:} \qquad \boldsymbol{\theta}_{t|t}^{\text{ex}} \;=\; \boldsymbol{\theta}_{t|t-1}^{\text{ex}} \;+\; \boldsymbol{H}_t\, \nabla \ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}^{\text{ex}}). \tag{48}$$

The "(one-step) filtered RMSE bound" is equivalent to that under Assumption 1a (misspecification) as both use the same implicit-gradient update (48):

$$\mathrm{RMSE}_{t|t} \leq \underbrace{\left(\sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} + \frac{\delta}{\mu_{\min}}\right)}_{=:\,c} \mathrm{RMSE}_{t|t-1} + \underbrace{\frac{\sigma_{\max}}{\mu_{\min}}}_{=:\,d}. \tag{49}$$

The "(one-step) predicted MSE bound" is equivalent to that using the ISD filter since both types of filters use prediction step (47): $\mathrm{MSE}_{t|t-1} \leq \|\mathbf{\Phi}^\star\|_2^2 \mathrm{MSE}_{t-1|t-1} + \sigma_\eta^2$. To obtain the "(one-step) predicted RMSE bound", we first take square roots on both sides and subsequently use that $\sqrt{z_1 + z_2} \leq \sqrt{z_1} + \sqrt{z_2}$ for non-negative scalars z_1 and z_2:

$$\text{RMSE}_{t|t-1} \leq \underbrace{\|\boldsymbol{\Phi}^\star\|_2}_{=:\, a} \text{RMSE}_{t-1|t-1} + \underbrace{\sigma_\eta}_{=:\, b} \tag{50}$$

Repeated substitution of the recursions (49) and (50) yields:

$$\begin{aligned} \text{RMSE}_{t|t} &\leq c^t a^{t-1} \text{RMSE}_{1|0} + d \sum_{i=0}^{t-1} (ca)^i + bc \sum_{i=0}^{t-2} (ca)^i \\ &= c^t a^{t-1} \text{RMSE}_{1|0} + d \frac{1-(ca)^t}{1-ca} + bc \frac{1-(ca)^{t-1}}{1-ca}, \quad ca \neq 1 \end{aligned}$$

where use a geometric series result: $\sum_{i=0}^{t} x^i = \left(1 - x^{t+1}\right)/(1-x)$ for $x \neq 1$. Moreover, under the following condition, which we refer to as the *contraction condition*:

$$\left(\sqrt{\frac{\mu_{\max} - 2\alpha}{\mu_{\min}}} + \frac{\delta}{\mu_{\min}} \right) \|\boldsymbol{\Phi}^\star\|_2 < 1 \tag{51}$$

The sums converge as $t \to \infty$, i.e. we can upper bound the *long-run root mean squared filtering error*:

$$\limsup_{t\to\infty} \text{RMSE}_{t|t} \leq \frac{\frac{\sigma_{\max}}{\sqrt{\mu_{\min}}} + \sigma_\eta \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}} \right)}{\sqrt{\mu_{\min}} - \left(\sqrt{\mu_{\max} - 2\alpha} + \frac{\beta}{\sqrt{\mu_{\min}}} \right) \|\boldsymbol{\Phi}^\star\|_2}. \tag{52}$$

D Proof Example 3

Here, we present an upper bound on the long-run MSE, and corresponding contraction condition for the *ISD* filter under Assumption 1b (correct specification), when the penalty matrix is a scalar multiple of the identity matrix. Suppose additionally Assumptions 1a and 3 hold. As $\boldsymbol{P}_t = \gamma \boldsymbol{I}_d$, this implies that $\mu_{\min} = \mu_{\max} = \gamma$. Then the contraction condition (44) can be simplified as follows: $\frac{\gamma}{\gamma + 2\alpha + \frac{\alpha^2}{\gamma}} \|\boldsymbol{\Phi}^\star\|_2^2 < 1 \Leftrightarrow \frac{\gamma^2}{(\gamma+\alpha)^2} \|\boldsymbol{\Phi}^\star\|_2^2 < 1$. Now, the contraction condition is *always* satisfied, since $\boldsymbol{\Phi}^\star = \boldsymbol{\Phi}^\dagger$ with $\rho(\boldsymbol{\Phi}^\dagger) \leq 1$ (Assumption

1b). For this choice of penalty matrix, the upper bound on the long-run MSE of the ISD filter (45) is minimal, and simplifies to:

$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\sigma_{\max}^2 + \sigma_\eta^2\gamma^2}{\alpha^2 + 2\alpha\gamma + (1 - \|\mathbf{\Phi}^\star\|_2^2)\gamma}. \tag{53}$$

Now, suppose the data is generated by a local level model: $y_t = \theta_t^\star + \varepsilon_t, \theta_{t+1}^\star = \theta_t^\star + \eta_t, \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2), \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2)$, where ε_t and η_t are mutually independent for all t, such that $\boldsymbol{P} = \gamma$.

Then $\sigma_{\max}^2 = \frac{1}{\sigma_\epsilon^2}$ and $\alpha = \frac{1}{\sigma_\epsilon^2}$. Substituting these in the RMSE bound (53), we obtain:

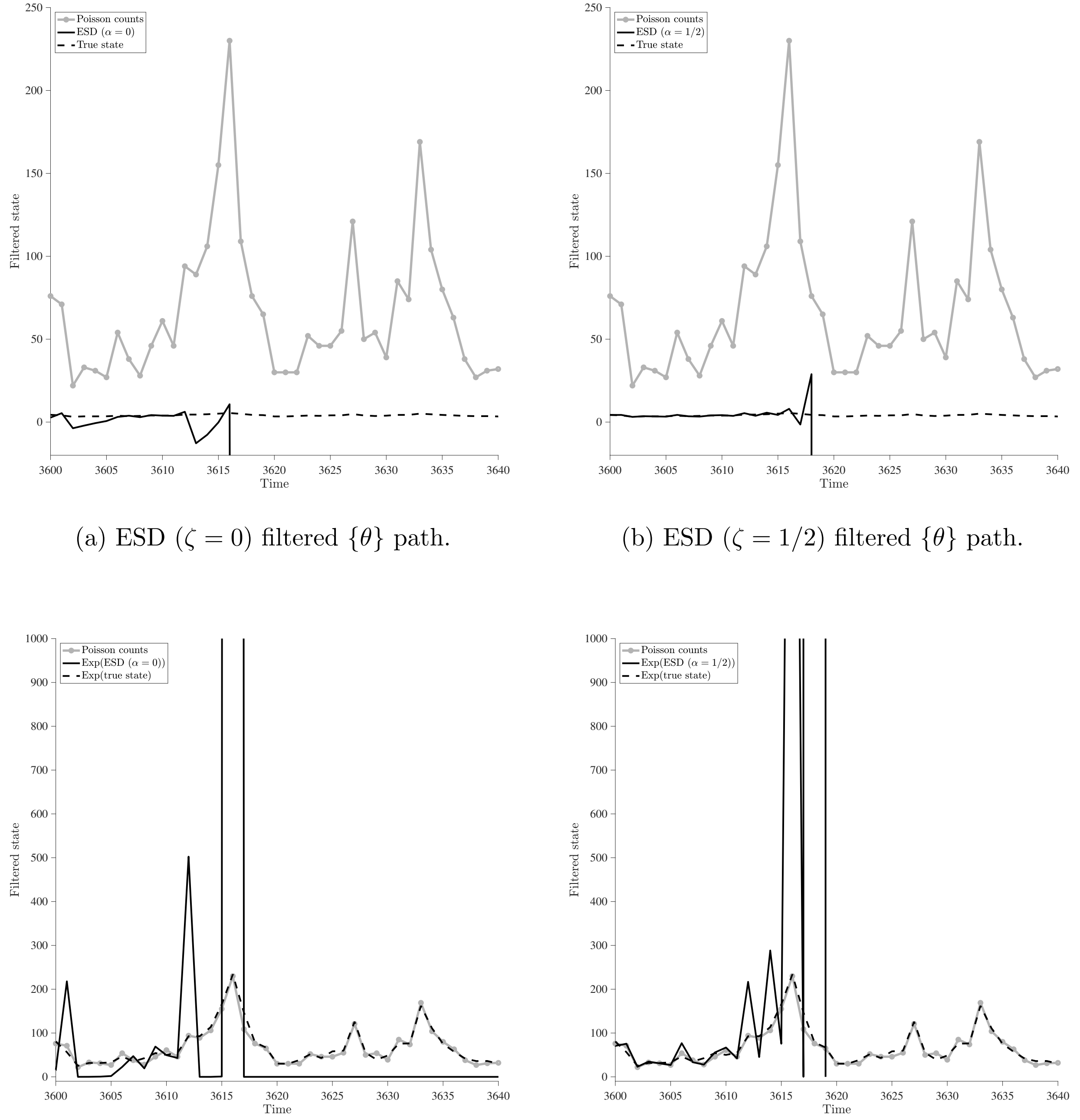
$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\sigma_\varepsilon^2 + \sigma_\eta^2\sigma_\varepsilon^4\gamma^2}{2\gamma\sigma_\varepsilon^2 + 1}. \tag{54}$$

Which is exactly minimized for a learning rate that is equal to the steady state Kalman covariance for the local level model.

For the RMSE bound and contraction condition of the *ESD* filter when $\boldsymbol{P}_t = \gamma\boldsymbol{I}_d$, we additionally assume $\beta-$smoothness of the observation log-density, i.e. Assumption 2b holds. The contraction condition (51) simplifies to: $(\sqrt{\frac{\gamma-2\alpha}{\gamma}} + \frac{\beta}{\gamma})\|\mathbf{\Phi}^\star\|_2 < 1$, which is *not* always satisfied, in contrast to the ISD filter. For this choice of penalty matrix, the upper bound on the long-run RMSE of the ESD filter (52) is minimal, and simplifies to:

$$\limsup_{t\to\infty} \mathrm{RMSE}_{t|t} \leq \frac{\sigma_{\max} + \sigma_\eta\left(\sqrt{\gamma^2 - 2\alpha\gamma} + \beta\right)}{\gamma - \left(\sqrt{\gamma^2 - 2\alpha\gamma} + \beta\right)\|\mathbf{\Phi}^\star\|_2}. \tag{55}$$

E Further simulation results



(a) ESD ($\zeta = 0$) filtered $\{\theta\}$ path.

(b) ESD ($\zeta = 1/2$) filtered $\{\theta\}$ path.

(c) ESD ($\zeta = 0$) filtered $\{\exp(\theta)\}$ path.

(d) ESD ($\zeta = 1/2$) filtered $\{\exp(\theta)\}$ path.

Figure 4: ESD = explicit score-driven. Zooming in on the filtered path of one true state $\{\theta_t^\dagger\}$ (cyan) based on the counts of a Poisson (grey dotted line) with state variation $\sigma_\eta = 0.39$ and state parameters $\omega^\dagger = -0.01$ and $\phi^\dagger = 0.98$ using an ESD filter with (a) identity scaling ($\zeta = 0$), and (b) inverse square root Fisher scaling ($\zeta = 1/2$). Figures (c) and (d) take instead the exponent of the true and filtered states.