TI 2025-006/III
Tinbergen Institute Discussion Paper

Stability and performance guarantees for misspecified multivariate score-driven filters

Simon Donker van Heel[1]
Rutger-Jan Lange[2]
Dick van Dijk[3]
Bram van Os[4]

1 Erasmus University Rotterdam

2 Erasmus University Rotterdam, Tinbergen Institute

2 Erasmus University Rotterdam, Tinbergen Institute

4 Vrije Universiteit Amsterdam

Stability and performance guarantees for misspecified multivariate score-driven filters*

Simon Donker van Heel†, Rutger-Jan Lange‡,
Dick van Dijk§ and Bram van Os¶

February 4, 2025

Abstract

We consider the problem of tracking latent time-varying parameter vectors under model misspecification. We analyze implicit and explicit score-driven (ISD and ESD) filters, which update a prediction of the parameters using the gradient of the logarithmic observation density (i.e., the score). In the ESD filter, the score is computed using the predicted parameter values, whereas in the ISD filter, the score is evaluated using the new, updated parameter values. For both filter types, we derive novel sufficient conditions for the exponential stability (i.e., invertibility) of the filtered parameter path and existence of a finite mean squared error (MSE) bound with respect to the pseudo-true parameter path. In addition, we present expressions for finite-sample and asymptotic MSE bounds. Our performance guarantees rely on mild moment conditions on the data-generating process, while our stability result is entirely agnostic about the true process. As a result, our primary conditions depend only on the characteristics of the filter; hence, they are verifiable in practice. Concavity of the postulated log density combined with simple parameter restrictions is sufficient (but not necessary) for ISD-filter stability, whereas ESD-filter stability additionally requires the score to be Lipschitz continuous. Extensive simulation studies validate our theoretical findings and demonstrate the enhanced stability and improved performance of ISD over ESD filters. An empirical application to U.S. Treasury-bill rates confirms the practical relevance of our contribution.

Keywords: Explicit and implicit-gradient methods; error bounds; pseudo-true parameters

*We thank the following individuals for their comments and suggestions: Janneke van Brummelen, Bernd Heidergott, Frank Kleibergen, André Lucas and the participants of the EIPC 2024, Tinbergen Institute PhD seminar 2024, Oxford Dynamic Econometrics Conference 2024, New York Camp Econometrics 2024, NESG 2024, ISEO Summer School 2024, IAAE 2024, and EEA-ESEM 2024. Remaining errors are our own.

†Econometric Institute, Erasmus University Rotterdam (corr. author, donkervanheel@ese.eur.nl)
‡Econometric Institute, Erasmus University Rotterdam (lange@ese.eur.nl)
§Econometric Institute, Erasmus University Rotterdam (djvandijk@ese.eur.nl)
¶Econometrics and Data Science Department, Vrije Universiteit Amsterdam (b.van.os2@vu.nl)

1 Introduction

Economic and financial variables often have characteristics that vary over time. One way of describing these fluctuations is through latent states, which can be tracked using filters that alternate between prediction and update steps. These two steps produce parameter estimates based on the lagged and contemporaneous information sets, respectively, as in Kalman's (1960) classic approach and its successors in the state-space literature (e.g., Durbin and Koopman, 2012).

In this paper, we consider multivariate score-driven filters. The update step in these filters moves the parameter estimate in the direction of the score, i.e., the gradient of the researcher-postulated logarithmic observation density. These filters may be misspecified in that the true state dynamics could be unknown and the postulated observation density incorrect. We differentiate between *explicit* and *implicit* score-driven (ESD and ISD) filters, which use the score evaluated in the researcher's predicted and updated parameters, respectively. Our theoretical contribution is twofold. First, for both types of filter we present sufficient conditions for *invertibility* (i.e., exponential stability) of the filtered parameter path (Theorem 1). These conditions relate only to the postulated observation density and are thus verifiable in practice. Second, under mild conditions on the true process, we establish performance guarantees for both filters (Theorem 2) by deriving upper bounds for the (non-)asymptotic mean squared errors (MSEs) of the filtered (or predicted) path relative to the pseudo-true parameter path.

We deviate in several ways from the literature on score-driven models (e.g., Artemova et al., 2022a, Harvey, 2022), which are also known as dynamic conditional score (DCS; Harvey, 2013) models and generalized autoregressive score (GAS; Creal et al., 2013) models. First, we do not assume that the true time-varying parameter follows score-driven dynamics. Rather, we take the more pessimistic view that our filter is score driven while, in general, the data-generating process (DGP) is not. Second, unlike standard score-driven approaches (but in line with the state-space literature), we differentiate between prediction and update steps. Separating these steps is conceptually useful and instrumental to our theory development. Third, we show that all score-driven filters in the literature (with the exception of Lange et al., 2024) are *explicit* score-driven filters. This is relevant, as our stability and performance guarantees in Theorems 1–2 are considerably stronger for implicit score-driven filters.

Theorem 1 guarantees filter invertibility, which is critical for the consistency of the maximum-likelihood estimator (MLE; Straumann and Mikosch, 2006). Although ESD filters are routinely calibrated using the MLE, a general invertibility result has only been

established for a scalar time-varying parameter; moreover, these results often rely on conditions involving the DGP (cf, Blasques et al., 2018, 2022). For ISD filters, the existing multivariate invertibility result in Lange et al. (2024) relies on the (assumed) concavity of the postulated logarithmic observation density. Theorem 1 presents novel sufficient conditions for the invertibility of multivariate ESD and ISD filters, where we remain agnostic with regard to the DGP and do not require log-concavity of the observation density.

Theorem 2 goes further by presenting sufficient conditions under which both filters are not only stable but also accurate. Analyzing tracking accuracy is important in machine learning (e.g., Cutler et al., 2023) and recently gained traction in the time-series literature (e.g., Brownlees and Llorens-Terrazas, 2024, Lange, 2024a). The best we can do under misspecification is to track the *pseudo*-true state. This means we cannot remain entirely agnostic about the DGP, but must impose some mild moment conditions; for example, that the pseudo-true state exists, while its increments have a finite variance. We also consider the more informative setting in which the DGP is a known state-space model with linear (but not necessarily Gaussian) state dynamics. In both settings, and for both ISD and ESD filters, Theorem 2 presents (non-)asymptotic performance guarantees by upper-bounding the MSE of the filtered and predicted parameter paths relative to the pseudo-true parameter path. The asymptotic MSE bounds can be minimized, often analytically, with respect to tuning parameters, such as the learning rate. In some special cases related to the Kalman filter, these minimized MSE bounds are exact (i.e., cannot be improved).

In both Theorems 1 and 2, ESD filters require more stringent conditions than ISD filters. Specifically, guarantees for the ESD filter are available only if the score is Lipschitz continuous in the parameter of interest. Assuming twice differentiability, this requires the Hessian of the postulated logarithmic density to be bounded. This Lipschitz condition is ubiquitous in the optimization literature (e.g., Nesterov, 2018), but appears to be neglected in the time-series literature (with the exception of Lange et al., 2024). While our theory suggests that this condition is sufficient for ESD filter stability, our simulation studies suggest that it is (near) necessary: when it does not hold and the underlying state is volatile, ESD filters may diverge, whereas ISD filters remain stable.

In three Monte Carlo experiments, we compare ISD and ESD filters in terms of stability, theoretical MSE bounds, and empirical MSEs. First, we investigate a high-dimensional setting based on Cutler et al. (2023), in which performance guarantees for both ISD and ESD filters are available. We compare both filters with three recent alternatives, finding that the ISD filter achieves the lowest MSE bounds and empirical MSEs for all time steps, state dimensions, and observation dimensions. The ISD filter is the only filter for which

the performance consistently improves as the Lipschitz constant of the score increases. Our MSE bounds are up to three orders of magnitude lower than those in Cutler et al. (2023). Second, we examine nine state-space models postulated in Koopman et al. (2016). While (finite) MSE bounds are available for all models when using an ISD filter, the same holds for the ESD filter only when the relevant Hessian is bounded, which occurs in just two models. When the underlying true parameter in the remaining seven models is volatile, the ESD filter is unstable. Third, we consider a Poisson count model, finding that the ISD filter remains stable when all investigated variants of the ESD filter are divergent.

Finally, we present an empirical application derived from Artemova et al. (2022b) involving U.S. Treasury-bill rate spreads. Here, we compare a new ISD version of a location model based on the Student's t distribution with its ESD version formulated in Harvey and Luati (2014). We show that the ISD filter yields a more favorable impact curve: it avoids the "zig-zagging" behavior of the filtered path when the data are constant, while enhancing the filter's responsiveness when the data are variable.

The remainder of this article is organized as follows. Section 2 outlines the features of ISD and ESD filters and links them with the standard literature. Sections 3 and 4 present our stability and performance guarantees, respectively. Sections 5 and 6 present our simulation studies and empirical illustration, respectively. Section 7 offers a brief conclusion. Finally, Appendix A contains proofs of theoretical results, Appendix B some further results in extended contexts.

2 Implicit and explicit score-driven filters

2.1 Problem setting

The $n \times 1$ observation $\boldsymbol{y}_t$ at times $t = 1, \dots, T$ is drawn from a true (but typically unknown) conditional observation density $p^0(\boldsymbol{y}_t|\boldsymbol{\vartheta}_t, \boldsymbol{\psi}^0, \mathcal{F}_{t-1})$. Here, $\boldsymbol{\vartheta}_t$ is a time-varying parameter vector taking continuous values in some parameter space $\boldsymbol{\Theta}^0$, $\boldsymbol{\psi}^0$ 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 $\boldsymbol{\psi}^0$ and $\mathcal{F}_{t-1}$ will be suppressed. In the case of discrete observations, $p^0(\boldsymbol{y}_t|\boldsymbol{\vartheta}_t)$ is interpreted as a probability rather than a density.

Our filters aim to track the true density $p^0(\boldsymbol{y}|\cdot)$ evaluated at the *true* parameter path $\{\boldsymbol{\vartheta}_t\}$ using some postulated density evaluated at some *filtered* parameter path. We denote by $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ the (misspecified) researcher-postulated density, where our model parameter $\boldsymbol{\theta}_t \in \boldsymbol{\Theta}$ denotes a vector of parameters contained in a (non-empty) convex parameter

space $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ that the researcher is interested in tracking (e.g., $\boldsymbol{\Theta}$ could be the positive quadrant). The dimension of $\boldsymbol{\theta}_t$ need not match that of the true parameter $\boldsymbol{\vartheta}_t$. Additional dependence on a static shape parameter $\boldsymbol{\psi}$ and/or other information available at time $t-1$ is permitted but suppressed; i.e., for conciseness, we write $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ as a short-hand for the more elaborate $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\psi}, \mathcal{F}_{t-1})$. The postulated density $p(\cdot|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ may implicitly include a *link function*; e.g., the parameter of interest could be the logarithmic variance, which is mapped to the positive real line by an exponential link function.

2.2 Filtering algorithm

Similar to Kalman's (1960) filter, our filtering algorithms alternate between update and prediction steps. As is standard, updated and predicted states reflect the researcher's estimates of the time-varying parameter a time t using the contemporaneous information set $\mathcal{F}_t$ and lagged information set $\mathcal{F}_{t-1}$, respectively. Updated (or "filtered") states are denoted $\{\boldsymbol{\theta}^j_{t|t}\}$, where the superscript $j \in \{\text{im}, \text{ex}\}$ indicates whether we consider *im*plicit score-driven (ISD) or *ex*plicit score-driven (ESD) filters. Sequences of predicted states are denoted as $\{\boldsymbol{\theta}^j_{t|t-1}\}$, again with $j \in \{\text{im}, \text{ex}\}$. ISD and ESD filters differ in their update steps, which employ either an implicit- or explicit-gradient method for all $t = 1, \ldots, T$:

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

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

where $\boldsymbol{H}_t \in \mathbb{R}^{k \times k}$ is the $\mathcal{F}_{t-1}$-measurable learning-rate matrix, assumed positive definite (i.e., $\boldsymbol{H}_t \succ \mathbf{O}_k$), $\nabla := \mathrm{d}/\mathrm{d}\boldsymbol{\theta}$ 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, explaining 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 often suppressed. The score on the right-hand side of (1) is evaluated at the point $\boldsymbol{\theta}^{\text{im}}_{t|t}$, which also appears on the left-hand side; this renders the method *implicit*. In contrast, the explicit update (2) is immediately computable, as the score is evaluated at $\boldsymbol{\theta}^{\text{ex}}_{t|t-1}$. We will see in Section (2.3) that implicit update (1) should be understood as a first-order condition associated with an optimization problem, while explicit update (2) solves a linearized version of the same problem.

Turning to the prediction step, both filters employ a linear first-order specification, i.e.,

$$\text{prediction step:} \qquad \boldsymbol{\theta}^j_{t|t-1} \;=\; (\boldsymbol{I}_k - \boldsymbol{\Phi})\, \boldsymbol{\omega} \;+\; \boldsymbol{\Phi}\, \boldsymbol{\theta}^j_{t-1|t-1}, \qquad j \in \{\text{im}, \text{ex}\}, \tag{3}$$

where $\boldsymbol{\omega}$ is a $k \times 1$ vector, $\boldsymbol{\Phi}$ is a $k \times k$ matrix, and $\boldsymbol{I}_k$ is the $k \times k$ identity. 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 $j \in \{\text{im}, \text{ex}\}$), when

convenient their superscripts are suppressed. The case $\boldsymbol{\Phi} = \boldsymbol{I}_k$ is permitted. Initializations $\boldsymbol{\theta}^j_{0|0} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ with $j \in \{\text{im}, \text{ex}\}$ are considered given. We emphasize that equations (1)–(3) specify the filtering algorithm, *not* the DGP. If $\boldsymbol{\Theta} \neq \mathbb{R}^k$, we can often ensure that prediction step (3) maps $\boldsymbol{\Theta}$ to itself by restricting $\boldsymbol{\omega}$ and $\boldsymbol{\Phi}$; e.g., if $\boldsymbol{\Theta}$ is the positive quadrant, taking $\boldsymbol{\omega} \in \boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ diagonal with $\mathbf{O}_k \preceq \boldsymbol{\Phi} \preceq \boldsymbol{I}_k$ suffices. Prediction step (3) could be generalized to allow for nonlinear and/or higher-order dynamics if these were found to be relevant. In economics and statistics, however, mean reversion is often important, while no (additional) information is available during the prediction step, such that a more complicated structure may not yield immediate benefits.

While ISD and ESD filters generally produce different outcomes, there is one famous case in which both methods, albeit with different learning rates, yield the same result.

Example 1 (Kalman's (1960) level update is an ISD ánd ESD update) *Consider the observation* $\boldsymbol{y}_t = \boldsymbol{d} + \boldsymbol{Z}\boldsymbol{\vartheta}_t + \boldsymbol{\varepsilon}_t$, *where* $\boldsymbol{y}_t, \boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{Z} \in \mathbb{R}^{n\times k}$, $\boldsymbol{\vartheta}_t \in \mathbb{R}^k$, *and* $\boldsymbol{\varepsilon}_t \sim \text{i.i.d.}\, \text{N}(\mathbf{0}_n, \boldsymbol{\Sigma}_\varepsilon)$ *with* $\boldsymbol{\Sigma}_\varepsilon \succ \mathbf{O}_n$. *Consider Kalman's predicted and updated covariance matrices* $\boldsymbol{P}^{\text{KF}}_{t|t-1}, \boldsymbol{P}^{\text{KF}}_{t|t} \succ \mathbf{O}_k$, *which are* $\mathcal{F}_{t-1}$ *measurable. If the postulated density* $p(\cdot|\boldsymbol{\theta})$ *matches the true density* $p^0(\cdot|\boldsymbol{\vartheta})$, *then Kalman's level update is (i) an ISD update (1) with learning-rate matrix* $\boldsymbol{H}^{\text{im}}_t = \boldsymbol{P}^{\text{KF}}_{t|t-1}$ *ánd (ii) an ESD update (2) with* $\boldsymbol{H}^{\text{ex}}_t = \boldsymbol{P}^{\text{KF}}_{t|t}$. *The implicit learning rate exceeds the explicit one, i.e.* $\boldsymbol{H}^{\text{im}}_t \succeq \boldsymbol{H}^{\text{ex}}_t$ *as* $\boldsymbol{P}^{\text{KF}}_{t|t-1} \succeq \boldsymbol{P}^{\text{KF}}_{t|t}$. *Kalman-filter generalizations can be based on its ISD or ESD representations as in Lange (2024a) or Ollivier (2018), respectively. For proof details, see Appendix A.1.*

2.3 Reformulation as optimization-based filters

Here we reformulate ISD update (1) as the following (multivariate) optimization problem:

$$\text{ISD update:} \qquad \boldsymbol{\theta}^{\text{im}}_{t|t} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmax}} \left\{ \ell\left(\boldsymbol{y}_t \mid \boldsymbol{\theta}\right) - \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{im}}_{t|t-1}\right\|^2_{\boldsymbol{P}_t} \right\}. \tag{4}$$

where $\boldsymbol{P}_t \in \mathbb{R}^{k\times k} \succ \mathbf{O}_k$ is the penalty matrix and $\|\cdot\|^2_{\boldsymbol{P}_t} = (\cdot)'\boldsymbol{P}_t(\cdot)$ is a squared (elliptic) Euclidean norm. Crucially, the penalty matrix is the inverse of the learning-rate matrix, i.e., $\boldsymbol{P}_t := \boldsymbol{H}_t^{-1} \succ \mathbf{O}_k$. The first-order condition associated with (4) reads $\mathbf{0}_k = \nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t}) - \boldsymbol{P}_t(\boldsymbol{\theta}^{\text{im}}_{t|t} - \boldsymbol{\theta}^{\text{im}}_{t|t-1})$. This can be re-arranged as $\boldsymbol{\theta}^{\text{im}}_{t|t} = \boldsymbol{\theta}^{\text{im}}_{t|t-1} + \boldsymbol{P}_t^{-1}\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t})$, which is exactly (1). First-order condition (1) can be viewed as a "short-hand" for the optimization (4); the latter clarifies that the ISD update maximizes the most recent log-likelihood contribution subject to a weighted quadratic penalty centered at the one-step-ahead prediction; for details, see Lange et al. (2024). In principle, the objective function could feature multiple stationary points, while the global solution—e.g., on the boundary of

the parameter space—is such that the first-order condition does not hold there. To rule out this possibility, Assumptions 1–3 in Section 3 will ensure the existence of a unique interior maximizer (4), which coincides with the (unique) solution to the first-order condition (1). Under these assumptions, formulations (1) and (4) are equivalent.

In optimization (4), we may take $\boldsymbol{\Theta} = \mathbb{R}^k$ to rule out boundary solutions and linearly approximate $\ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ around the prediction to recover the ESD update (2):

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

The associated first-order condition reads $\mathbf{0}_k = \nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{ex}}_{t|t-1}) - \boldsymbol{P}_t(\boldsymbol{\theta}^{\text{ex}}_{t|t} - \boldsymbol{\theta}^{\text{ex}}_{t|t-1})$. This can be rearranged as $\boldsymbol{\theta}^{\text{ex}}_{t|t} = \boldsymbol{\theta}^{\text{ex}}_{t|t-1} + \boldsymbol{P}_t^{-1} \nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{ex}}_{t|t-1})$, which is exactly (2). The linear approximation in (5) could potentially be improved by using a quadratic approximation, but to the best of our knowledge, in the filtering literature this has never been tried.

The full optimization (4) has some advantages over its linearized version (5). First, optimization (4) guarantees $\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t}) \geq \ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t-1})$; i.e., the update always improves the fit of the postulated logarithmic density. The same does not hold for its linearized version (5); for details, see Appendix B.1. Second, linearization (5) leading to first-order condition (2) implies that $\boldsymbol{\theta}^{\text{ex}}_{t|t}$ must be allowed to take values in all of $\mathbb{R}^k$; i.e. no boundaries can be imposed. While link functions can be incorporated in the definition of $p(\boldsymbol{y}|\boldsymbol{\theta})$, we will see that the additional non-linearity may complicate the filter's theoretical properties (e.g., by affecting Lipschitz constants).

2.4 Static (hyper-)parameter estimation

Both filters involve (hyper-)parameters that must be estimated, including the learning rate matrices $\{\boldsymbol{H}^j_t\}$ in the parameter update steps (1)–(2), the parameters $\boldsymbol{\omega}^j$ and $\boldsymbol{\Phi}^j$ in the prediction steps (3), and the shape parameters $\boldsymbol{\psi}^j$ (if any) in the postulated observation density, where $j \in \{\text{ex}, \text{im}\}$. In the numerical examples presented in Sections 5 and 6, we use a static learning rate matrix, i.e., $\boldsymbol{H}^j_t = \boldsymbol{H}^j \succ \mathbf{O}_k, \forall t \geq 1$ and $j \in \{\text{im}, \text{ex}\}$. We collect the static parameters in a single column vector; $\boldsymbol{\upsilon}^j := (\text{vech}(\boldsymbol{H}^j)', \boldsymbol{\omega}^{j'}, \text{vec}(\boldsymbol{\Phi}^j)', \boldsymbol{\psi}^{j'})'$, where $\text{vech}(\cdot)$ and $\text{vec}(\cdot)$ denote the half-vectorization and vectorization matrix operations, respectively. The static parameter $\boldsymbol{\upsilon}^j$ is calibrated by maximum-likelihood (ML) estimation based on the prediction-error decomposition as is the *de facto* standard in observation-driven modeling (e.g., Creal et al., 2013, eq. 15), i.e.

$$\hat{\boldsymbol{\upsilon}}^j := \underset{\boldsymbol{\upsilon}^j\in\boldsymbol{\Upsilon}}{\text{argmax}} \sum_{t=1}^{T} \log p\left(\boldsymbol{y}_t \mid \boldsymbol{\theta}^j_{t|t-1}, \boldsymbol{\psi}^j, \mathcal{F}_{t-1}\right), \qquad j \in \{\text{im}, \text{ex}\}, \tag{6}$$

where the right-hand side depends on $\boldsymbol{\upsilon}^j$ directly (through the shape parameters $\boldsymbol{\psi}^j$) and indirectly (through the predicted parameter path $\{\boldsymbol{\theta}^j_{t|t-1}\}$). The optimization domain $\boldsymbol{\Upsilon}$ is the subset of the parameter space for which the learning rate is positive definite ($\boldsymbol{H}^j \succ \mathbf{O}_k$) and $\boldsymbol{\psi}^j$ lies in its allowable domain. As $T \to \infty$, our estimate $\hat{\boldsymbol{\upsilon}}^j$ does not necessarily approach a "true" parameter; this is because our postulated density may be misspecified and the DGP need not be score driven. Instead, the limiting version of $\hat{\boldsymbol{\upsilon}}^j$ can be interpreted as a "pseudo-true" static parameter that minimizes the Kullback-Leibler divergence from our misspecified filter to the true process (e.g. Blasques et al., 2022, Thm 4.6). For maximum-likelihood procedure (6) to be well defined, the filter must at least be *invertible* (e.g., Bougerol, 1993 and Straumann and Mikosch, 2006). Invertibility for univariate ESD filters is shown in Blasques et al. (2022); however, estimation is routinely performed for multivariate filters. Section 3 presents sufficient conditions for multivariate ISD and ESD filter invertibility without requiring any knowledge of the DGP.

2.5 Comparison with standard (explicit) score-driven filters

Typical score-driven models (e.g., Artemova et al., 2022a Harvey, 2022) are equivalent to—what we would call—*explicit* score-driven filters. To demonstrate this, we substitute the ESD update (2) into the prediction (3) to obtain the following prediction-to-prediction recursion that is common in score-driven modeling:

$$\boldsymbol{\theta}^{\text{ex}}_{t+1|t} = (\boldsymbol{I}_k - \boldsymbol{\Phi})\boldsymbol{\omega} + \boldsymbol{\Phi}\,\boldsymbol{\theta}^{\text{ex}}_{t|t-1} + \boldsymbol{\Phi}\,\boldsymbol{H}_t\,\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{ex}}_{t|t-1}). \tag{7}$$

Some freedom remains in choosing the learning-rate matrix $\boldsymbol{H}_t$; throughout this article, we ensure $\boldsymbol{H}_t \succ \mathbf{O}_k$. The literature on score-driven models (e.g., Creal et al., 2013, eq. 4–5) typically takes the matrix in front of the score to be $\boldsymbol{A}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\text{ex}}_{t|t-1})^{-\zeta}$ for some arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$. Here, $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) := \int p(\boldsymbol{y}|\boldsymbol{\theta})\nabla\ell(\boldsymbol{y}|\boldsymbol{\theta})\nabla'\ell(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y}$ is the postulated Fisher information matrix, while $\zeta \in \{0, 1/2, 1\}$ determines the scaling. An important downside of $\boldsymbol{A}$ being (allowed to be) arbitrary is that the method may fail to select a direction of ascent; for further details, see Appendix B.2.

3 Invertibility of score-driven filters

Before we can give performance guarantees of ISD and ESD filters, a first step is to demonstrate the (exponential) stability of both filters (e.g., Guo and Ljung, 1995), famously known in the econometrics literature as *filter invertibility* (e.g., Straumann and Mikosch, 2006). This property, which is important in its own right as well as a necessary ingredient

for the maximum-likelihood estimation of static parameters (e.g., Blasques et al., 2022), means that filtered paths originating from different starting points (but based on identical data) converge exponentially fast over time. The general invertibility result in Blasques et al., 2022 for the ESD filter applies to the univariate setting (i.e., for a scalar time-varying parameter), while the conditions guaranteeing invertibility are tied to the DGP. A multivariate invertibility result is available for the ISD filter, but only as long as the postulated density is log-concave (Lange et al., 2024). This section presents a framework for demonstrating invertibility for both ISD or ESD filters that (i) requires no knowledge of the DGP, (ii) allows for a multivariate parameter vector, and (iii) does not impose log-concavity on the researcher-postulated density.

3.1 Assumptions related to the filter

To prove the stability of the update and prediction steps, here we make assumptions only about the postulated density $p(\cdot|\boldsymbol{\theta}_t)$, the autoregressive matrix $\boldsymbol{\Phi}$, and the penalty matrices $\{\boldsymbol{P}_t\}$, which we will assume to be constant over time (i.e., $\boldsymbol{P}_t = \boldsymbol{P}, \forall t$). As these are all under the researcher's control, our assumptions have the advantage of being practically verifiable. Our first condition relates to the curvature of the postulated logarithmic density, where (bounded) violations of concavity are permitted.

Assumption 1 (Negative postulated Hessian is bounded below) *The Hessian matrix* $\boldsymbol{\mathcal{H}}(\boldsymbol{y}, \boldsymbol{\theta}) := \nabla^2 \log p(\boldsymbol{y} \mid \boldsymbol{\theta})$ *corresponding to the* postulated *density is well-defined for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and all* $\boldsymbol{y} \in \mathbb{R}^n$. *Moreover, the negative Hessian is lower and upper bounded as* $\alpha \boldsymbol{I}_k \preceq -\boldsymbol{\mathcal{H}}(\boldsymbol{y}, \boldsymbol{\theta}) \preceq \beta \boldsymbol{I}_k$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and all* $\boldsymbol{y} \in \mathbb{R}^n$, *where* $\beta \in (0, \infty]$ *could be infinite, while* $\alpha (\leq \beta)$ *is bounded away from negative infinity.*

Assumption 1 is useful because it allows us to cover a variety of relevant cases. If the postulated logarithmic density is twice differentiable and concave, then $\alpha \geq 0$ such that the negative Hessian is positive semi-definite. If the logarithmic density is strongly concave, then $\alpha > 0$, where α is the parameter of strong concavity. If it (locally) fails to be concave, we have $\alpha < 0$. We can write α as its positive part minus its negative part, i.e. $\alpha = \alpha^+ - \alpha^-$, where $\alpha^+ := \max\{0, \alpha\}$ and $\alpha^- := \max\{0, -\alpha\}$. Then α^- can be viewed as the maximum *violation* of concavity, which in Assumption 1 is assumed to be bounded (i.e. $\alpha^- < \infty$). If the gradient is Lipschitz continuous, then $\beta < \infty$, in which case the Lipschitz constant is $L := \max\{\alpha^-, \beta\}$. This reasoning demonstrates that α^- and β can be viewed as one-sided Lipschitz constants. Stability for the ISD filter will depend only the value of α (thus allowing $\beta = \infty$), while for the ESD filter, both α and β will be relevant. Under

the combination of concavity ($0 \leq \alpha \leq \beta$) and a Lipschitz-continuous gradient ($\beta < \infty$), the Lipschitz constant is $L = \max\{0, \beta\} = \beta$.

We will analyze optimization (4) via its first-order condition (1). Hence our next assumptions relate to the existence and uniqueness of a solution to this stationary condition. Assumption 2 postulates that optimization (4) is proper strongly concave. This condition is automatically satisfied if $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}|\boldsymbol{\theta})$ is bounded above, while the penalty term "overwhelms" any concavity violation by the postulated logarithmic density; this can be achieved by taking $\lambda_{\min}(\boldsymbol{P}) > \alpha^-$. If the parameter space is unbounded (e.g., $\boldsymbol{\Theta} = \mathbb{R}^k$), Assumption 2 automatically implies the existence and uniqueness of a solution to first-order condition (1). For a general parameter space $\boldsymbol{\Theta} \neq \mathbb{R}^k$, however, there is one small caveat in that the first-order condition may not be satisfied unless we additionally rule out boundary solutions. For this reason, Assumption 3 separately requires the argmax (4) to lie in the *interior* of $\boldsymbol{\Theta}$; to maintain consistency over time, we also require the prediction step (3) to map $\boldsymbol{\Theta}$ to itself. This situation presents no loss of generality compared with the ESD filter, which in any case requires $\boldsymbol{\Theta} = \mathbb{R}^k$.

Assumption 2 (Optimization problem (4) is proper strongly concave) *The penalty matrix is static and positive definite, i.e.* $\boldsymbol{P}_t = \boldsymbol{P} \succ \mathbf{O}_k, \forall t$. *Moreover, the objective function in optimization problem (4) is proper strongly concave; in the context of Assumption 1, this implies that* $\lambda_{\min}(\boldsymbol{P}) > \alpha^-$.

Assumption 3 (Optimization problem (4) has an interior solution) *For the ESD filter (2)–(3), we impose* $\boldsymbol{\Theta} = \mathbb{R}^k$. *For the ISD filter (3)–(4), we allow* $\boldsymbol{\Theta} \neq \mathbb{R}^k$ *but impose that the argmax (4) lies in the* interior *of* $\boldsymbol{\Theta}$ *for any observation* $\boldsymbol{y}_t \in \mathbb{R}^n$ *and any prediction* $\boldsymbol{\theta}^{im}_{t|t-1} \in \boldsymbol{\Theta}$, *while (*$\boldsymbol{\omega}$ *and* $\boldsymbol{\Phi}$ *are such that) the prediction step (3) maps* $\boldsymbol{\Theta}$ *to itself.*

Remark 1 *For simplicity, we take Assumptions 1–3 to hold* for all $\boldsymbol{y} \in \mathbb{R}^n$. *As they relate only to the postulated density, this has the advantage that we remain entirely* agnostic *with regard to the DGP. We could allow a mild dependence on the DGP by taking Assumption 1–3 to hold "almost surely in* $\boldsymbol{y}_t$*" instead of "for all* $\boldsymbol{y}_t$*". Lemma 1 and Theorem 1 (to follow) would then hold almost surely in the data* $\{\boldsymbol{y}_t\}$, *rather than for all data* $\{\boldsymbol{y}_t\}$.

3.2 Stability guarantees

Assumptions 1–3 are sufficient to derive the following update step stability result.

Lemma 1 (Update stability) *Let Assumptions 1–3 hold. Fix $t \geq 1$ and let two predictions $\boldsymbol{\theta}^{j}_{t|t-1} \in \boldsymbol{\Theta}$ with $j \in \{\text{im}, \text{ex}\}$ be given. Based on these predictions, compute ISD and ESD updates (1) and (2), yielding $\boldsymbol{\theta}^{\text{im}}_{t|t}$ and $\boldsymbol{\theta}^{\text{ex}}_{t|t}$, respectively. Then, for all $\boldsymbol{y}_t$,*

$$\left\| \frac{\mathrm{d}\boldsymbol{\theta}^{\text{im}}_{t|t}}{\mathrm{d}\boldsymbol{\theta}^{\text{im}\,\prime}_{t|t-1}} \right\|_{\boldsymbol{P}} \leq 1 - \frac{\alpha^{+}}{\lambda_{\max}(\boldsymbol{P}) + \alpha^{+}} + \frac{\alpha^{-}}{\lambda_{\min}(\boldsymbol{P}) - \alpha^{-}}, \tag{8}$$

$$\left\| \frac{\mathrm{d}\boldsymbol{\theta}^{\text{ex}}_{t|t}}{\mathrm{d}\boldsymbol{\theta}^{\text{ex}\,\prime}_{t|t-1}} \right\|_{\boldsymbol{P}} \leq 1 - \min\left\{ \frac{\alpha^{+}}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^{-}}{\lambda_{\min}(\boldsymbol{P})}, 2 - \frac{\beta}{\lambda_{\min}(\boldsymbol{P})} \right\}. \tag{9}$$

To the best of our knowledge, inequalities (8)–(9) are new. Despite Lemma 1's apparent simplicity, its proof is somewhat involved and quite lengthy. Lemma 1 considers the sensitivity of the updated parameter with respect to the predicted parameter, as measured by the Jacobian matrix, in the $\boldsymbol{P}$-weighted matrix norm. We call the updating step stable (in the $\boldsymbol{P}$-norm) if the weighted norm does not exceed unity. By equation (8), for the implicit update to be stable for all observations $\boldsymbol{y}_t$ and all predictions, it is both necessary and sufficient to have $\alpha \geq 0$, which can be achieved by postulating a (sufficiently smooth) log-concave observation density. In particular, the update is always contractive if $\alpha > 0$, while it may be expansive if $\alpha < 0$. The largest possible expansion is bounded due to Assumption 2 (i.e. $\lambda_{\min}(\boldsymbol{P}) > \alpha^{-}$). An expansion in the update step may be permitted as long as the contraction in the prediction step is sufficiently strong (see Lemma 2 below).

As inequality (9) shows, ESD update stability additionally requires $\beta < \infty$; otherwise the right-hand side would be unbounded. While $\alpha \geq 0$ remains necessary for the ESD update to be stable, it is no longer sufficient: if β is large such that the second argument of $\min\{\cdot, \cdot\}$ dominates, we also need $1 - (2 - \beta/\lambda_{\min}(\boldsymbol{P})) \leq 1$. Using $1/\lambda_{\min}(\boldsymbol{P}) = \lambda_{\max}(\boldsymbol{H})$, this additional requirement can be succinctly written as $\lambda_{\max}(\boldsymbol{H}) \leq 2/\beta$. Hence the largest eigenvalue of the learning-rate matrix should be bounded above by $2/\beta$, which we recognize as a classic condition in the optimization literature (e.g., Nesterov, 2003, p. 80). A notable implication of this condition is that the learning-rate matrix must shrink to zero as $\beta \to \infty$. Hence ESD update stability cannot be guaranteed unless the gradient is Lipschitz continuous; this observation, which in the time-series literature is apparently new, will turn out to have serious consequences (see discussion after Theorem 1).

Remark 2 *An important novelty of Lemma 1 is its use of the* $\boldsymbol{P}$-weighted *(i.e., elliptic) matrix norm. A similar result would not generally be possible when using the Euclidean norm. This dependence on using the $\boldsymbol{P}$-norm may explain why multivariate invertibility results for ESD filters are, hitherto, in short supply. Using the $\boldsymbol{P}$-norm is possible in the current article only because we separate the update step (which depends on $\boldsymbol{P}$) from the prediction step (which depends on $\boldsymbol{\omega}$ and $\boldsymbol{\Phi}$). Using the $\boldsymbol{P}$-norm is no exception in the (static)*

optimization literature (e.g. Duchi, 2018, Thm. 4.3.5). The standard formulation of score-driven models, however, employs a single prediction-to-prediction recursion parametrized in terms of $\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{H} = \boldsymbol{\Phi}\boldsymbol{P}^{-1}$*, where* $\boldsymbol{A}$ *is often taken to be unrestricted. This parametrization in terms of* $\boldsymbol{A}$ *has prevented researchers from considering the* $\boldsymbol{P}$*-norm. As all norms are equivalent, the choice of norm is ultimately one of convenience, allowing us to prove invertibility (see Theorem 1 to follow) in both the* $\boldsymbol{P}$*-weighted and Euclidean norm.*

To prove stability of the filtered path over time, it is sufficient to show that the composition of each prediction step with each update step is stable. Having established the stability of the update step in the $\boldsymbol{P}$ norm, the next result does the same for the linear first-order prediction step (3).

Lemma 2 (Prediction stability) *Fix* $\boldsymbol{\Phi} \in \mathbb{R}^{k\times k}$ *and* $\boldsymbol{P} \succ \boldsymbol{O}_k$*. Then*

$$\|\boldsymbol{\Phi}\|^2_{\boldsymbol{P}} \leq 1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}, \tag{10}$$

where $\gamma := \lambda_{\min}(\boldsymbol{P} - \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi}) \in \mathbb{R}$*,* $\gamma^+ := \max\{0, \gamma\}$ *and* $\gamma^- := \max\{0, -\gamma\}$*.*

To simplify, in the scalar case (i.e., $\boldsymbol{\Phi} = \phi$ and $\boldsymbol{P} = \rho > 0$) we have $\gamma = \rho(1 - \phi^2)$. Then Lemma 2 says that if γ is positive (negative), then the prediction step is contractive (expansive) in the $\boldsymbol{P}$ norm. In this case, $\gamma \geq 0 \Leftrightarrow \phi^2 \leq 1$, which corresponds with the usual notion of (univariate) stability. In the multidimensional case, the scalar $\gamma \in \mathbb{R}$ controls how much the weighted norm $\|\boldsymbol{\Phi}\|^2_{\boldsymbol{P}}$ lies below unity (if $\gamma > 0$) or above unity (if $\gamma < 0$). As $\boldsymbol{P}$ and $\boldsymbol{\Phi}$ are part of the filter, $\gamma := \lambda_{\min}(\boldsymbol{P} - \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi})$ is computable in practice. Of course, we could ensure $\gamma \geq 0$ by imposing $\boldsymbol{P} \succeq \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi}$, which demonstrates a link with the discrete-time Lyapunov equation. This inequality shows that the unit-root case $\boldsymbol{\Phi} = \boldsymbol{I}_k$ is permitted (in which case $\gamma = 0$). While $\gamma \geq 0$ is sufficient for prediction stability, we may allow $\gamma < 0$ if the contraction in the update step (see Lemma 1) is sufficiently strong. Lemma 2 could be avoided if we computed $\|\boldsymbol{\Phi}\|^2_{\boldsymbol{P}}$ directly (i.e., without using the bound (10)).

Lemmas 1–2 combine to yield our first main result regarding the invertibility of multivariate ISD and ESD filters, which are misspecified in the sense that we have made no assumptions on the DGP.

Theorem 1 (Invertibility of misspecified multivariate score-driven filters) *Let Assumptions 1–3 hold. Consider two starting points* $\boldsymbol{\theta}_{0|0}$ *and* $\widetilde{\boldsymbol{\theta}}_{0|0}$*, and corresponding filtered paths, denoted* $\{\boldsymbol{\theta}^j_{t|t}\}$ *and* $\{\widetilde{\boldsymbol{\theta}}^j_{t|t}\}$*, respectively, based on the same data* $\{\boldsymbol{y}_t\}$ *and using either the ISD (*$j = \text{im}$*) or ESD (*$j = \text{ex}$*) filter. Then, for all data sequences* $\{\boldsymbol{y}_t\}$*,*

$$\|\boldsymbol{\theta}^j_{t|t} - \widetilde{\boldsymbol{\theta}}^j_{t|t}\|^2_{\boldsymbol{P}} \quad \leq \quad \tau^t_j \ \|\boldsymbol{\theta}_{0|0} - \widetilde{\boldsymbol{\theta}}_{0|0}\|^2_{\boldsymbol{P}}\,, \qquad \forall t \geq 0, \qquad j \in \{\text{im}, \text{ex}\},$$

where the contraction coefficients τ_j for $j \in \{\text{im}, \text{ex}\}$ are defined as

$$\tau_{\text{im}} := \left(1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}\right)\left(1 - \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P}) + \alpha^+} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P}) - \alpha^-}\right)^2, \tag{11}$$

$$\tau_{\text{ex}} := \left(1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}\right)\left(1 - \min\left\{\frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}, \frac{2\lambda_{\min}(\boldsymbol{P}) - \beta}{\lambda_{\min}(\boldsymbol{P})}\right\}\right)^2, \tag{12}$$

where γ is defined in Lemma 2. Under the (sufficient) condition $\tau_j < 1$ for $j \in \{\text{im}, \text{ex}\}$, $\lim_{t\to\infty} \|\boldsymbol{\theta}^j_{t|t} - \widetilde{\boldsymbol{\theta}}^j_{t|t}\| = 0$ for any starting points $\boldsymbol{\theta}_{0|0}$ and $\widetilde{\boldsymbol{\theta}}_{0|0}$ and any data sequence $\{\boldsymbol{y}_t\}$; moreover, this convergence to zero is exponentially fast.

To the best of our knowledge, Theorem 1 provides the first invertibility result for a multivariate ESD filter with conditions that are both verifiable and independent of the (unknown) DGP (cf. Blasques et al., 2022). We deviate from other approaches in that we require neither fat-tailed log-likelihood functions nor uniformly bounded third derivatives (cf. Harvey, 2013 and Harvey and Luati, 2014). Theorem 1 also provides the first invertibility result for the ISD filter allowing non-concavity of the postulated log-density (i.e., allowing $\alpha < 0$) and a non-contractive prediction step (i.e., allowing $\gamma \leq 0$). Hence our analysis covers the unit-root case $\boldsymbol{\Phi} = \boldsymbol{I}_k$ (for which $\gamma = 0$). An important advantage of Theorem 1 is that it relies only on properties of the *postulated* (as opposed to the true) observation density along with the (researcher-specified) autoregressive matrix $\boldsymbol{\Phi}$ and the penalty matrix $\boldsymbol{P}$; hence, conditions (11)–(12) are verifiable in practice.

For the ISD filter, inequality (11) indicates that it is sufficient to have both $\gamma \geq 0$ and $\alpha \geq 0$, with at least one of these inequalities being strict. In this case, both multiplicative terms in (11) are bounded above by unity, while at least one is strictly below one. The same logic dictates that either $\gamma < 0$ (due to an expansive prediction step) or $\alpha < 0$ (due to non-concavity) could be permitted (but not both), as long as (11) holds.

For the ESD filter, equation (12) highlights that to formulate a sufficient condition for invertibility, we additionally require the Lipschitz condition $\beta < \infty$. To understand this condition, note that the sensitivity of the ESD output (which involves the score) with respect to the input is governed by the Hessian. To guarantee stability, the Hessian should be bounded, which implies $\beta < \infty$. While this Lipschitz condition is ubiquitous in the (static) optimization literature (e.g., Nesterov, 2018), it appears to be virtually unknown in the time-series literature. Indeed, with the exception of Lange et al. (2024), it is, to the best of our knowledge, never mentioned as being being necessary or desirable in the large literature on ESD filters. This is problematic because a lack of gradient Lipschitz continuity may cause ESD filter divergence. That this fact has gone unnoticed in the time-series literature may be due to authors subconsciously focusing on cases in which the

performance of ESD filters is unproblematic. As we will see in Section 5, if the Lipschitz condition $\beta < \infty$ fails and the ESD filter is misspecified in the sense that the DGP is a state-space model, then we can typically make the ESD filter diverge by increasing the variability of the state process. The critical level of the state variability at which divergence starts to occur is, beforehand, unknown. In contrast, as (11) is independent of β, ISD filters remain stable for all (i.e., even large) state variabilities.

Theorem 1 further deviates from the literature on score-driven models in that it does not assume that the DGP is score driven; in fact, we made no assumptions on the DGP whatsoever. If we were to follow this ubiquitous assumption (i.e., taking the filter to be the DGP), then Theorem 1 would automatically imply that the filtered path converges to the true path—at least as long as we can consistently estimate the static parameters (including $\boldsymbol{\omega}, \boldsymbol{\Phi}, \boldsymbol{P}$), as would be guaranteed under certain regularity conditions (see Blasques et al., 2022 and Blasques et al., 2018, Remark 3.1). Under consistent parameter estimation, therefore, Theorem 1 would guarantee that the true path is perfectly recovered. This is consistent with Koopman et al. (2016, Table 5), who find that (explicit) score-driven filters are particularly accurate when the DGP is also score driven. Here we take the more pessimistic view that our filter is score driven while the DGP, in general, is not. The DGP could be almost entirely unknown or given by some state-space model. The next section investigates what theoretical performance guarantees can be given for score-driven filters in this misspecified setting.

As Example 2 illustrates, Theorem 1 generalizes the invertibility result for ESD filters in Blasques et al. (2018) and Blasques et al. (2022) to the multivariate setting.

Example 2 (Student's t scalar location model) *Blasques et al. (2022, sec. 5) consider a scalar local-level model, originally formulated in Harvey and Luati (2014), where the postulated observation density is a Student's t distribution and the ESD filtering dynamics are based on the score scaled by $\sigma^2/(1+\nu^{-1})$. For the logarithmic Student's t density multiplied by $\sigma^2/(1+\nu^{-1})$ (to match the above literature), Assumption 1 is satisfied with $\alpha = -1/8$ (log-concavity is violated) and $\beta = 1$ (the gradient is Lipschitz). This model features a (scalar) mean-reversion parameter $\boldsymbol{\Phi} = \phi$ and a (scalar) penalty parameter $\boldsymbol{P} = \rho > 0$. Assumption 2 requires $\rho > 1/8$, i.e. the learning rate $\eta = 1/\rho$ is positive and bounded above by 8. (Here we deviate from Blasques et al. (2022), who permit negative learning rates.) According to Theorem 1, a sufficient condition for invertibitily is $|\phi|\ (1 - \min\{-\eta/8, 2-\eta\})^2 < 1$, or, equivalently,*

$$|\phi| \left(\max\left\{1 + \frac{\eta}{8}, \eta - 1\right\}\right)^2 < 1. \tag{13}$$

Combining this condition with $\eta < 8$, it is sufficient if we have (i) $|\phi| < 1$ along with (ii) $\eta < \min\{8, 8(1/|\phi| - 1), 1 + 1/|\phi|\}$. This result is similar to (but slightly different from) Blasques et al. (2022), who use a different parametrisation in treating the product of (our) ϕ and η as the learning-rate parameter. According to Theorem 1, a sufficient condition for invertibitily of the ISD filter is

$$|\phi| < (1 + \alpha\eta)^2 = \left(1 - \frac{\eta}{8}\right)^2. \tag{14}$$

Similarly, η must be close to zero when $|\phi|$ is close to unity.

4 Performance guarantees for score-driven filters

This section investigates to what extent we are able to track the true density $p^0(\boldsymbol{y}_t|\boldsymbol{\vartheta}_t)$; hence, we must reintroduce some consideration of the true process. Even so, we will continue to consider the multivariate and misspecified case; that is, we permit $k > 1$ and do not assume that the DGP follows score-driven dynamics. To measure performance, we consider the *weighted mean squared filtering error* $\text{MSE}^{\boldsymbol{W}}_{t|t} := \mathbb{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{W}}]$ and the *weighted mean squared prediction error* $\text{MSE}^{\boldsymbol{W}}_{t|t-1} := \mathbb{E}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{W}}]$, where $\boldsymbol{W} \succ \mathbf{O}_k$ is a positive definite weight matrix and $\boldsymbol{\theta}^\star_t$ denotes the pseudo-true parameter defined below. Similar to the previous section (where we used a weighted norm), the introduction of a *weighted* MSE allows some freedom even if we are ultimately interested in usual case, for which $\boldsymbol{W} = \boldsymbol{I}_k$. While the true parameter $\boldsymbol{\vartheta}_t$ and pseudo-true parameter $\boldsymbol{\theta}^\star_t$ in general have different dimensions, they coincide whenever the postulated density equals the true density (i.e., when $p(\boldsymbol{y}_t|\cdot) = p^0(\boldsymbol{y}_t|\cdot)$).

Definition 1 (Pseudo-true parameter) *Consider a true distribution $p^0(\boldsymbol{y}_t|\boldsymbol{\vartheta}_t)$ modeled by some postulated distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$. Then $\boldsymbol{\theta}^\star_t := \text{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int p^0(\boldsymbol{y}|\boldsymbol{\vartheta}_t)\ell(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y}$ is the pseudo-true parameter, provided a unique solution exists.*

4.1 Assumptions related to the DGP

To allow the pseudo-true parameter path $\{\boldsymbol{\theta}^\star_t\}$ to be tracked over time with some level of accuracy, we require two moment conditions. First, we need the increments of the pseudo-true process to have finite covariance matrix. Second, we need the the score evaluated at the pseudo-truth to have a bounded unconditional second moment.

Assumption 4 (Unknown DGP with moment conditions) *Consider a true distribution $p^0(\boldsymbol{y}_t|\boldsymbol{\vartheta}_t)$ modeled by some postulated distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$. Assume for $t = 1, \dots, T$ that:*

(i) **(Drift)** *The pseudo-truth* $\boldsymbol{\theta}_t^\star$ *exists and its increments have finite second (cross) moments. That is,* $\operatorname{cov}(\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t-1}^\star) \preceq \boldsymbol{Q}$, *where* $\mathbf{O}_k \preceq \boldsymbol{Q} \in \mathbb{R}^{k\times k}$ *with* $q^2 := \operatorname{tr}(\boldsymbol{Q}) < \infty$.

(ii) **(Score)** *The first moment of the postulated score, evaluated at the pseudo-truth, is zero, while its second moment is bounded. That is,* $\int p^0(\boldsymbol{y}|\boldsymbol{\vartheta}_t)\nabla\ell(\boldsymbol{y}|\boldsymbol{\theta}_t^\star)\,\mathrm{d}\boldsymbol{y} = \mathbf{0}_k$ *and* $\sigma_t^2 := \mathbb{E}[\|\nabla\ell(\boldsymbol{y}|\boldsymbol{\theta}_t^\star)\|^2] \leq \sigma^2 < \infty$.

Assumption 4(i), which posits that the pseudo-true increments $\{\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t-1}^\star\}$ have a finite second moment uniformly across time, is unrelated to our particular filtering setup. It is needed by any method that requires an MSE-type loss to be computed. In theory, Assumption 4(i) requires the existence of a number of moments on the increments of the true process $\{\boldsymbol{\vartheta}_t - \boldsymbol{\vartheta}_{t-1}\}$, where the precise number depends on the severity of the level of misspecification of $p(\boldsymbol{y}_t|\cdot)$ relative to $p^0(\boldsymbol{y}_t|\cdot)$. If the mapping from the true to the pseudo-true parameter is Lipschitz continuous—which trivially includes the case $p(\boldsymbol{y}_t|\cdot) = p^0(\boldsymbol{y}_t|\cdot)$—then bounding the variance of true increments suffices. Assumption 4(i) is substantially weaker than common assumptions in the literature, which often imposes that the pseudo-true parameter increments are *uniformly* bounded (e.g., Wilson et al., 2019; Cao et al., 2019; Simonetto et al., 2020; Lanconelli and Lauria, 2024; Bernardi et al., 2024). In our notation, this would imply the existence of a positive constant C such that $\|\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star\| < C$, $\forall t \geq 1$. Although ubiquitous, this assumption excludes many important DGPs, such as the linear Gaussian state-space model relevant for the Kalman filter.

Assumption 4(ii), which is standard in the stochastic-approximation literature (e.g., Toulis and Airoldi, 2017; Bottou et al., 2018), requires the second moment (under the true density) of the postulated score evaluated in the pseudo truth to be bounded. This condition relaxes the common requirement that the second moment of the score is bounded in the *entire* space $\boldsymbol{\Theta}$ (e.g., Lehmann and Casella, 1998; Nemirovski et al., 2009). When the observation density matches the true density (i.e., $p(\boldsymbol{y}_t|\cdot) = p^0(\boldsymbol{y}_t|\cdot)$), Assumption 4(ii) implies that the maximum eigenvalue of the Fisher information matrix is bounded.

Naturally, our performance guarantees can be further improved if the DPG is known. To investigate this setting in some level of generality, Assumption 5 states that the DGP is a known state-space model with an general observation density but linear (and possibly non-Gaussian) state dynamics. While more specific than Assumption 4, Assumption 5 is still more general than the setting of linear Gaussian state-space models that are used to derive Kalman's (1960) filter. Indeed, this classic setting falls under Assumption 5 if we additionally impose that the observation density $p^0(\cdot|\boldsymbol{\vartheta}_t)$ is Gaussian, while its the mean is a linear transformation of $\boldsymbol{\vartheta}_t$ and the state increments $\{\boldsymbol{\xi}_t\}$ are i.i.d. Gaussian with a known

covariance matrix. Assumption 5 implies Assumption 4(i) (with $\boldsymbol{Q} = \boldsymbol{\Sigma}_\xi$ and $q = \sigma_\xi$) but not Assumption 4(ii); hence, Assumption 5 will be invoked jointly with Assumption 4.

Assumption 5 (Known DGP: State-space model with linear state dynamics) *For all $t = 1, \ldots, T$, let the true density $p^0(\cdot|\cdot)$ be known, while the true parameter $\{\boldsymbol{\vartheta}_t\}$ follows first-order linear dynamics:*

$$\boldsymbol{\vartheta}_t = (\boldsymbol{I}_k - \boldsymbol{\Phi}_0)\,\boldsymbol{\omega}_0 + \boldsymbol{\Phi}_0 \boldsymbol{\vartheta}_{t-1} + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim \text{i.i.d.}\,(\boldsymbol{0}, \boldsymbol{\Sigma}_\xi), \tag{15}$$

with finite covariance matrix $\mathbf{O}_k \preceq \boldsymbol{\Sigma}_\xi \in \mathbb{R}^{k\times k}$ i.e. $\sigma_\xi^2 := \operatorname{tr}(\boldsymbol{\Sigma}_\xi) < \infty$. The intercept parameter vector $\boldsymbol{\omega}_0 \in \mathbb{R}^d$ and auto-regressive matrix $\boldsymbol{\Phi}_0 \in \mathbb{R}^{k\times k}$ with spectral radius $\varrho(\boldsymbol{\Phi}_0) < 1$ are assumed known.

When Assumption 5 holds, we allow our algorithms to take advantage of this additional information. That is, we use $p(\boldsymbol{y}_t|\cdot) = p^0(\boldsymbol{y}_t|\cdot)$ in our updates (1) and (2) for the ISD and ESD filters, respectively. Similarly, in the prediction step (3), we use the true state-transition parameters, i.e. we set $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}_0$. This procedure is similar that for the Kalman filter, which also uses the true observation density to compute updates, and the true state-transition parameters to compute predictions. Because the true density is known, the pseudo-true and true parameters coincide, i.e. $\boldsymbol{\theta}_t^\star = \boldsymbol{\vartheta}_t$, for $t = 1, \ldots, T$. This case is as close to "correct specification" as we will get; however, because the DGP is a (parameter-driven) state-space model, our score-driven filters are *still* misspecified. In general, therefore, score-driven filters may be suboptimal; the only exception is when the state-space model is linear and Gaussian, as in Example 1, in which case our score-driven filters with appropriately tuned learning-rate matrices $\boldsymbol{H}^j$ for $j \in \{\text{im}, \text{ex}\}$ match the steady-state Kalman filter.

4.2 Performance guarantees

In order to provide a concise overview of our performance guarantees, we lay out a general approach that accommodates ISD or ESD filters in the setting of either Assumption 4 (alone) or Assumptions 4–5 (together). Analyzing both filters under two different sets of assumptions yields four possible cases. In each case, we will derive linear system of two inequalities involving $\{\text{MSE}^{\boldsymbol{P}}_{t|t}\}$ and $\{\text{MSE}^{\boldsymbol{P}}_{t|t-1}\}$ as follows:

$$\text{updating-error bound}: \quad \text{MSE}^{\boldsymbol{P}}_{t|t} \leq a\ \text{MSE}^{\boldsymbol{P}}_{t|t-1} + \underbrace{b}_{\text{noise}} \tag{16}$$

$$\text{prediction-error bound}: \quad \text{MSE}^{\boldsymbol{P}}_{t|t-1} \leq c\ \text{MSE}^{\boldsymbol{P}}_{t-1|t-1} + \underbrace{d}_{\text{drift}} \tag{17}$$

for all $t \geq 1$. Crucially, the values $a, b, c, d \geq 0$ are specific to each case. The values of a and b depend on which update is used (ISD or ESD), where a governs the contraction rate (ideally, $a < 1$), while b is related to the noisiness of the score as in Assumption 4(ii). The values of c and d depend only on whether Assumption 4 holds (alone) or, in addition, Assumption 5. Here, c is related to the contraction in the prediction step via the auto-regressive matrix $\boldsymbol{\Phi}$ (ideally, $c < 1$), while d is related to the "drifting" (i.e., stochastic increments) of the pseudo-true state as specified in Assumption 4(i). Similar to the stability section, we use a $\boldsymbol{P}$-weighted MSE for convenience: concavity (and, for the ESD update, a Lipschitz-continuous gradient) will be sufficient to obtain a non-expansive update (i.e., $a \leq 1$). Even as we use the $\boldsymbol{P}$-weighted MSE for our computation, we can can always translate back to the usual (Euclidean) MSE.

By a standard geometric-series result, the pair (16)–(17) yields the following *finite-sample error bound* if $a\,c \neq 1$:

$$\mathrm{MSE}_{t|t}^{\boldsymbol{P}} \quad \leq \quad \underbrace{a^t c^{t-1}\,\mathrm{MSE}_{1|0}^{\boldsymbol{P}}}_{\text{initialisation error}} \; + \; \underbrace{\frac{1-a^t c^t}{1-a\,c}\,b}_{\text{noise error}} \; + \; \underbrace{\frac{1-a^{t-1}c^{t-1}}{1-a\,c}\,a\,d}_{\text{drift error}}, \qquad t \geq 1, \tag{18}$$

where three contributions to the weighted MSE can be identified: (i) the initial error $\mathrm{MSE}_{1|0}^{\boldsymbol{P}}$, (ii) the noisiness of the observations (related to b), and (iii) the drifting of the underlying state (related to d). Under the condition $a\,c < 1$, by letting $t \to \infty$ we obtain the following *asymptotic error bounds*:

$$\limsup_{t\to\infty}\ \mathrm{MSE}_{t|t}^{\boldsymbol{P}} \leq \frac{b\,+\,a\,d}{1-a\,c}, \qquad \limsup_{t\to\infty}\ \mathrm{MSE}_{t|t-1}^{\boldsymbol{P}} \leq c\frac{b\,+\,a\,d}{1-a\,c} + d = \frac{b\,c + d}{1-ac}. \tag{19}$$

The first bound, which relates to updates, follows directly from (18). The second bound, which relates to predictions, follows by combining the first bound with inequality 17. Again, the values of $a, b, c, d \geq 0$ are specific to each of the four cases. In both asymptotic bounds, the effect of the initialization has vanished, while the noise and drift terms remain. Both bounds are linearly increasing in the noise b, which is logical as a higher b means that observations are less informative about $\boldsymbol{\theta}_t^\star$. Similarly, both bounds grows linearly with the drift d, reflecting that a more volatile state $\{\boldsymbol{\theta}_t^\star\}$ is harder to track. Furthermore, the bound is increasing in $a\,c$, which can be interpreted as the joint contraction rate of the update and prediction steps. To see that the $\boldsymbol{P}$-weighted MSE implies a bound on the usual MSE, note that $\mathrm{MSE}_{t|t}^{\boldsymbol{W}} \leq \lambda_{\max}(\boldsymbol{W})\mathrm{MSE}_{t|t} \leq (\lambda_{\max}(\boldsymbol{W})/\lambda_{\min}(\boldsymbol{P}))\mathrm{MSE}_{t|t}^{\boldsymbol{P}}$ for any $\boldsymbol{W} \succ \mathbf{O}_k$. These inequalities are equalities if $\boldsymbol{W}$ and $\boldsymbol{P}$ are scalar multiples of the identity matrix. For $\boldsymbol{W} = \boldsymbol{I}_k$, we find $\mathrm{MSE}_{t|t} \leq 1/(\lambda_{\min}(\boldsymbol{P}))\mathrm{MSE}_{t|t}^{\boldsymbol{P}}$.

Theorem 2 uses our assumptions on the filter (Assumptions 1–3) together with the assumptions on the DGP (Assumptions 4–5) to derive specific values of $a, c, b, d \geq 0$ that

deliver finite (non-)asymptotic error bounds (18)–(19) for the ISD and ESD filters.

Theorem 2 (MSE bounds for misspecified multivariate score-driven filters) *Let Assumptions 1–4 hold and Assumption 5 only when indicated. In addition, assume that the Hessian $\boldsymbol{\mathcal{H}}(\boldsymbol{y}_t, \boldsymbol{\theta})$ is Riemann integrable in $\boldsymbol{\theta}$ with probability one in $\boldsymbol{y}_t$. Then, $\forall \chi, \epsilon > 0$, the values of a, b, c, d in the pair (16)–(17) for the ISD and ESD filters are:*

Update	Contractive factor	Noise
ISD (1)	$a = \left(1 - \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})+\alpha^+} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})-\alpha^-}\right)^2$	$b = \frac{a\sigma^2}{\lambda_{\min}(\boldsymbol{P})}$
ESD (2)	$a = \left(1 - \min\left\{\frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}, \frac{2\lambda_{\min}(\boldsymbol{P})-\beta}{\lambda_{\min}(\boldsymbol{P})}\right\}\right)^2 + \frac{\chi^2 L^2}{\lambda_{\min}(\boldsymbol{P})^2}$	$b = \frac{(1+1/\chi^2)\sigma^2}{\lambda_{\min}(\boldsymbol{P})}$
Prediction	Multiplicative factor	Drift
A4	$c = (1+\epsilon^2)\left(1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}\right)$	$d = \lambda_{\max}(\boldsymbol{P})(1+\frac{1}{\epsilon^2})(\\|\boldsymbol{I}_k - \boldsymbol{\Phi}\\|_2\, s_\omega + q)^2$
A4–5	$c = 1 - \frac{\gamma_0^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma_0^-}{\lambda_{\min}(\boldsymbol{P})}$	$d = \lambda_{\max}(\boldsymbol{P})\sigma_\xi^2$

where $\gamma_0^+ := \max\{0, \gamma_0\}$, $\gamma_0^- := \max\{0, -\gamma_0\}$ and $\gamma_0 := \lambda_{\min}(\boldsymbol{P} - \boldsymbol{\Phi}_0{}'\boldsymbol{P}\boldsymbol{\Phi}_0)$. When Assumption 5 holds, $ac < 1$ is sufficient for both filters to obtain finite MSE bounds. When Assumption 5 does not hold, we must ensure $d < \infty$ by imposing either (i) $s_\omega^2 := \sup_t \mathbb{E}[\\|\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}\\|^2] < \infty$ and/or (ii) $\boldsymbol{\Phi} = \boldsymbol{I}_k$. In these cases, the finite sample and asymptotic MSE bounds are obtained by substituting a, b, c, d in (18) and (19), respectively.

Theorem 2 expresses a, b, c, d in terms of other quantities defined in Assumptions 1–5. In addition, there are two "free" parameters, denoted by $\epsilon > 0$ and $\chi > 0$, which derive from the use of Young's inequality. These can be chosen freely by the researcher such that Theorem 2 presents a *family* of admissible bounds (indexed by these two parameters). Choosing ϵ and χ amounts to a trade-off: smaller values lead to lower multiplicative constants a and c but larger additive constants b and d. The bound for the ISD filter contains no free parameters when Assumption 5 holds. When Assumption 5 does not hold, it contains one free parameter, $\epsilon > 0$, which can be analytically optimized to yield the best (i.e., smallest) possible bound; see Corollary 1 below. In practice, any free parameters can also be chosen by numerically minimizing the bound with respect to these parameters.

Corollary 1 (Optimal Young's parameter for the ISD filter) *Let Assumptions 1–4 hold and assume that sufficient condition (11) for stability in Theorem 1 is satisfied, implying $\tau_{\text{im}} = ac/(1+\epsilon^2) < 1$ for a given $\epsilon > 0$. In addition, let (i) $s_\omega^2 := \sup_t \mathbb{E}[\\|\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}\\|^2] < \infty$ and/or (ii) $\boldsymbol{\Phi} = \boldsymbol{I}_k$. Then, for $\tau_{\text{im}} > 0$, the value of the Young's parameter $\epsilon^2 > 0$ that minimizes $\limsup_{t\to\infty} \text{MSE}_{t|t}^{\boldsymbol{P}}$ is given by*

$$\epsilon_\star^2 = \frac{1 - \tau_{\text{im}}}{\tau_{\text{im}} + \sqrt{\tau_{\text{im}} + \tau_{\text{im}}(1-\tau_{\text{im}})\frac{\sigma^2}{\lambda_{\max}(\boldsymbol{P})\lambda_{\min}(\boldsymbol{P})(\\|\boldsymbol{I}_k - \boldsymbol{\Phi}\\|_2\, s_\omega + q)^2}}} < \infty. \tag{20}$$

In Theorem 2, we find for both filters that ac closely resembles the contraction coefficients τ_{im} and τ_{ex} given in conditions (11)–(12) for invertibility in Theorem 1; in fact, if Assumption 5 holds we have for the ISD filter that $ac = \tau_{\text{im}}$, whereby $\gamma = \gamma_0$. In the three other cases, we have that as Young's parameters ϵ and χ tend to 0 that ac tends to τ_{im} or τ_{ex} for the ISD and ESD filters, respectively. Because Young's parameters can be chosen arbitrarily close to 0, we have that $\tau_{\text{im}}, \tau_{\text{ex}} < 1$ implies the existence of $\epsilon, \chi > 0$ such that $ac < 1$. Effectively, this implies that under the conditions of Theorem 2, we have that the invertibility conditions (11) for ISD or (12) for ESD from Theorem 1 are sufficient to obtain finite asymptotic MSE bounds.

Similar to the discussion of Theorem 1, we have that concavity (i.e. $\alpha \geq 0$) is sufficient for non-expansiveness of the ISD update. In this case, the ISD contraction rate $a \leq 1$ presented in Theorem 2 is better than in Lange (2024a, Eq. 20), while matching Lange et al. (2024, Corol. 1). For the non-concave case (i.e. $\alpha < 0$), the expression is new. Intuitively, concavity ensures that the score evaluated in any prediction on average (i.e., in expectation over $\boldsymbol{y}$) points in the direction of the pseudo-truth $\boldsymbol{\theta}_t^\star$, while its magnitude grows sufficiently fast as the prediction $\boldsymbol{\theta}_{t|t-1}^j$, $j \in \{\text{im}, \text{ex}\}$ is moved away from $\boldsymbol{\theta}_t^\star$. Crucially, the ESD update requires not only $\alpha > 0$ to obtain a contraction but also a global maximum curvature, i.e. $\beta \leq L < \infty$, to prevent excessive updates in the direction of the observation $\boldsymbol{y}_t$. The same requirement was needed to prove stability of the ESD filter in Theorem 1. Furthermore, we note that the penalty matrix $\boldsymbol{P}$ for both filters is a key determinant of the contractive factor a and the noise b and presents a trade-off. That is, higher values of $\boldsymbol{P}$ (as measured by its maximum and minimum eigenvalues) reflect a more persistent filter with a closer to 1, but lower b. This generates a trade-off that can be optimized; see also Lange et al. (2024, Corollary 1). If the true process is a known state-space model (Assumption 5 holds) and the penalty matrix is a scalar multiple of the identity matrix ($\boldsymbol{P} = \rho \boldsymbol{I}_k$ for some $\rho > 0$), then a solution can easily be derived analytically; Corollary 2 treats the ISD case.

Corollary 2 (Optimal penalty parameter for the ISD filter) *Let Assumptions 1–5 hold. Furthermore, assume that the log-likelihood contribution is strongly concave ($\alpha > 0$), that the penalty matrix is a scalar multiple of the identity matrix, i.e. $\boldsymbol{P} = \rho \boldsymbol{I}_k$ for some $\rho > 0$, and that $\sigma, \sigma_\xi > 0$ such that the noise b and drift d are non-zero. Then, the value of $\rho > 0$ that minimizes $\limsup_{t\to\infty} \text{MSE}_{t|t}$ is*

$$\rho_\star = \frac{\sigma^2 \left(1 - \|\boldsymbol{\Phi}_0\|^2\right) - \alpha^2 \sigma_\xi^2 + \sqrt{\left(\alpha^2 \sigma_\xi^2 - \sigma^2 \left(1 - \|\boldsymbol{\Phi}_0\|^2\right)\right)^2 + 4\alpha^2 \sigma^2 \sigma_\xi^2}}{2\alpha \sigma_\xi^2} < \infty. \tag{21}$$

Note that $\lim_{\sigma \downarrow 0} \rho_\star = 0$, reflecting that we should completely neglect the prediction $\boldsymbol{\theta}_{t|t-1}$ and fully rely on the data $\boldsymbol{y}_t$ if the latter's information about $\boldsymbol{\vartheta}_t$ is without noise.

In the bottom half of the table in Theorem 2, we see that when only Assumption 4 (but not Assumption 5) holds, we need some additional conditions to ensure $d < \infty$. First, if the pseudo-true process has a finite unconditional variance, such that $s_\omega^2 := \sup_t \mathbb{E}[\|\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}\|^2]$ is finite, then we can use a mean-reverting prediction step (i.e., $\boldsymbol{\Phi} \neq \boldsymbol{I}_k$). In this case, the bound is improved if s_ω is minimized, which means we should ideally set $\boldsymbol{\omega}$ to be the "correct" average level of the pseudo-truth, i.e. take $\boldsymbol{\omega} = \mathbb{E}[\boldsymbol{\theta}_t^\star]$. Hence, if the pseudo-true process has a finite variance, then we can use a mean-reverting prediction step; moreover, ideally the filter should mean-revert to the correct unconditional level. With regard to choosing $\boldsymbol{\Phi}$, we see that $\boldsymbol{\Phi}$ closer to $\mathbf{O}_k$ (as measured by the norm $\|\boldsymbol{\Phi}\|_2^2$) produces a stronger contraction via c but a larger drift d, which leads to a trade-off that can be optimized.

Second, if the variance of the pseudo-true process is unbounded (e.g. because the process is integrated of order one), then we can *still* obtain a finite MSE simply by setting $\boldsymbol{\Phi} = \boldsymbol{I}_k$, in which case $\boldsymbol{\omega}$ becomes irrelevant. Hence we can even track unit-root dynamics as long as $\boldsymbol{\Phi} = \boldsymbol{I}_k$. When Assumption 5 holds, such that the DGP is a known state-space model and the filter uses the *true* state-transition parameters (i.e. $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}_0$), then d is automatically finite as $\sigma_\xi^2 < \infty$ due to Assumption 4.

While Assumption 5 is not essential, it allows for better performance guarantees. In some cases, our bounds may even be exact. Example 3 illustrates this by considering the same setting as in Example 1 (i.e., a linear Gaussian state-space model). Specializing to the univariate case, we see that the asymptotic bound implied by Theorem 2 cannot be further improved, as it matches the asymptotic MSE implied by the Kalman filter.

Example 3 (Exact MSE bounds of ISD filter for AR(1) plus noise model) *Consider an AR(1) model with additive observation noise, defined by the observation equation* $y_t = \vartheta_t + \varepsilon_t$, *where* $\varepsilon_t \sim$ i.i.d. $\mathrm{N}(0, \sigma_\varepsilon^2)$, *and linear first-order state dynamics* $\vartheta_{t+1} = (1 - \phi_0)\omega_0 + \phi_0 \vartheta_t + \xi_t$, *where* $\xi_t \sim$ i.i.d. $\mathrm{N}(0, \sigma_\xi^2)$. *Let the learning rate be positive and constant,* $\boldsymbol{H}_t = \boldsymbol{P}_t^{-1} = \rho^{-1} = \eta > 0$, *for all* t. *As in Kalman's classic setting, we assume that Assumption 5 holds; i.e., the observation density and parameters* ω_0 *and* ϕ_0 *are known to the researcher. Then the asymptotic MSE bound*

$$\limsup_{t \to \infty} \mathrm{MSE}_{t|t} \leq \frac{\eta^2 \sigma_\varepsilon^2 + \sigma_\xi^2 \sigma_\varepsilon^4}{2\eta \sigma_\varepsilon^2 + \eta^2 + (1 - \phi_0^2)\sigma_\varepsilon^4}, \tag{22}$$

is minimized when the learning rate η *is chosen as*

$$\eta_\star = \frac{\sigma_\xi^2 - \sigma_\varepsilon^2 (1 - \phi_0^2) + \sqrt{\sigma_\xi^4 + \sigma_\varepsilon^4 (1 - \phi_0^2)^2 + 2\sigma_\xi^2 \sigma_\varepsilon^2 (1 + \phi_0^2)}}{2}. \tag{23}$$

For the local-level model, for which $\phi_0 = 1$, the minimizer $\eta_\star$ simplifies to

$$\eta_\star = \frac{\sigma_\xi^2 + \sqrt{\sigma_\xi^4 + 4\sigma_\xi^2\sigma_\varepsilon^2}}{2} = \frac{\sigma_\varepsilon^2}{2}\left(\frac{\sigma_\xi^2}{\sigma_\varepsilon^2} + \sqrt{\frac{\sigma_\xi^4}{\sigma_\varepsilon^4} + 4\frac{\sigma_\xi^2}{\sigma_\varepsilon^2}}\right). \tag{24}$$

As it turns out, $\eta_\star$ equals the steady-state variance of prediction errors in the Kalman filter as given in Durbin and Koopman (2012, p. 37) or Harvey (1989, p. 175), where $\sigma_\xi^2/\sigma_\varepsilon^2$ is the signal-to-noise ratio. Substituting the minimizer $\eta_\star$ back into the ISD update, we obtain the exact form of Kalman's level update, i.e., with the (optimal) Kalman gain. Since the Kalman filter is optimal for this model, our minimized MSE bounds are exact. The derivation of the optimal learning rate $\eta_\star$ is given in Appendix A.8.

5 Simulation studies

We present three simulation studies comparing ISD and ESD filters in terms of stability and filtering performance:

1. Section 5.1 investigates a high-dimensional setting from Cutler et al. (2023). In addition to our two filters, we include three recent competitors: an online version of Nesterov's method (ONM, Nesterov, 2018) and two stochastic gradient methods, proposed in Madden et al. (2021) and Cutler et al. (2023). The ISD filter achieves the lowest empirical MSE of all five filters for all time steps, state dimensions, observation dimensions, and almost all Lipschitz gradient constants. The ISD filter is the only filter for which the performance consistently improves as the Lipschitz constant of the score increases; moreover, the associated MSE bounds always lower and sometimes much lower (e.g., by a factor $\sim 10^3$) than those in Cutler et al. (2023).

2. Section 5.2 investigates nine DGPs from Koopman et al. (2016). For the ISD filter, MSE bounds are available for all nine models. For the ESD filter, in contrast, only two models have (finite) MSE bounds. For remaining seven, the ESD filter turns out to be unstable when the underlying state volatility is increased.

3. Section 5.3 investigates a state-space Poisson count model, finding that the ISD filter remains stable, while all investigated variants of the ESD filter are divergent, even when the observation density is correctly specified.

5.1 Least-squares recovery as in Cutler et al. (2023)

Observation equation. Following Madden et al. (2021, p. 453–454) and Cutler et al. (2023, p. 41–42), we investigate a high-dimensional linear model. Conditional on the latent state $\boldsymbol{\vartheta}_t \in \mathbb{R}^k$, each observation $\boldsymbol{y}_t \in \mathbb{R}^n$ is independently drawn from a Gaussian distribution $\mathrm{N}(\boldsymbol{A}\boldsymbol{\vartheta}_t, \boldsymbol{\Sigma})$. The (static) matrix $\boldsymbol{A} \in \mathbb{R}^{n\times k}$ is generated using Haar-distributed orthogonal matrices to ensure that it has rank $k \leq n$, while its singular values are equally spaced in the interval $[\sqrt{\alpha}, \sqrt{\beta}\,]$ with $0 < \alpha \leq \beta < \infty$. As we will see, this matches the notation in Assumption 1. The covariance matrix $\boldsymbol{\Sigma} = \sigma^2/(n\beta)\boldsymbol{I}_n$ with $0 < \sigma^2 < \infty$ is formulated so that $\sigma^2 < \infty$ matches the notation in Assumption 4(ii).

State transition. The latent state $\boldsymbol{\vartheta}_t \in \mathbb{R}^k$ evolves according to a random walk $\boldsymbol{\vartheta}_{t+1} = \boldsymbol{\vartheta}_t + \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t$ is i.i.d. but non-Gaussian, being drawn uniformly from the surface area of a k-dimensional sphere of radius $\sigma_\xi > 0$, which matches the notation in Assumption 5. The initial state $\boldsymbol{\vartheta}_0$ is drawn from the surface of a sphere with radius 10.

Postulated density. Following Cutler et al. (2023), we are interested in least-squares recovery, which means that all directions are equally important. Hence we postulate a Gaussian density in which the covariance matrix is the identity, i.e.

$$\ell(\boldsymbol{y}_t|\boldsymbol{\theta}) = -\frac{1}{2}\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}_t\|^2 - \frac{k}{2}\log(2\pi). \tag{25}$$

The Hessian with respect to $\boldsymbol{\theta}$ is constant at $\boldsymbol{\mathcal{H}} = -\boldsymbol{A}'\boldsymbol{A}$. Due to the construction of $\boldsymbol{A}$, the eigenvalues of $-\boldsymbol{\mathcal{H}}$ lie in $[\alpha, \beta]$ with $0 < \alpha \leq \beta < \infty$, matching the notation in Assumption 1. Even as the observation log-density (25) is technically misspecified, the pseudo-true parameter still equals the true parameter (i.e., $\boldsymbol{\vartheta}_t = \boldsymbol{\theta}^\star_t, \forall t$). Moreover, taking the postulated covariance matrix to be a *scalar multiple* of the identity would lead to identical updates, just with a different learning rate. Parameter σ^2 matches the notation in Assumption 4(ii), as $\mathbb{E}\|\nabla(\boldsymbol{y}_t|\boldsymbol{\vartheta}_t)\|^2 = \mathbb{E}\|\boldsymbol{A}'(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{\vartheta}_t)\|^2 = \mathrm{tr}(\boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A}) = \beta\,\mathrm{tr}(\boldsymbol{\Sigma}) = \sigma^2 < \infty$.

Filter specification. Our filters are initialized in the origin and employ the identity mapping as their prediction steps; i.e., $\boldsymbol{\theta}^j_{0|0} = \boldsymbol{0}_k$ and $\boldsymbol{\theta}^j_{t+1|t} = \boldsymbol{\theta}^j_{t|t}$ with $j \in \{\mathrm{ex}, \mathrm{im}\}$. Following Cutler et al. (2023), we consider learning-rate matrices that are scalar multiples of the identity. The ESD update with $\boldsymbol{H}^{\mathrm{ex}}_t = \eta^{\mathrm{ex}}\boldsymbol{I}_k$ and $\eta^{\mathrm{ex}} > 0$ then equals

$$\text{ESD update:} \qquad \boldsymbol{\theta}^{\mathrm{ex}}_{t|t} = \boldsymbol{\theta}^{\mathrm{ex}}_{t|t-1} + \eta^{\mathrm{ex}}\boldsymbol{A}'(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{\theta}^{\mathrm{ex}}_{t|t-1}),$$

where $\nabla(\boldsymbol{y}_t|\boldsymbol{\theta}^{\mathrm{ex}}_{t|t-1}) = \boldsymbol{A}'(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{\theta}^{\mathrm{ex}}_{t|t-1})$ is the (explicit) score. For the ISD update, the first-order condition $\boldsymbol{\theta}^{\mathrm{im}}_{t|t} = \boldsymbol{\theta}^{\mathrm{im}}_{t|t-1} + \eta^{\mathrm{im}}\boldsymbol{A}'(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{\theta}^{\mathrm{im}}_{t|t})$ can be solved for $\boldsymbol{\theta}^{\mathrm{im}}_{t|t}$ to yield

$$\begin{aligned}\text{ISD update:} \qquad \boldsymbol{\theta}^{\mathrm{im}}_{t|t} &= \left(\boldsymbol{I}_k + \eta^{\mathrm{im}}\boldsymbol{A}'\boldsymbol{A}\right)^{-1}\left(\boldsymbol{\theta}^{\mathrm{im}}_{t|t-1} + \eta^{\mathrm{im}}\boldsymbol{A}'\boldsymbol{y}_t\right), \\ &= \boldsymbol{\theta}^{\mathrm{im}}_{t|t-1} + \eta^{\mathrm{im}}\left(\boldsymbol{I}_k + \eta^{\mathrm{im}}\boldsymbol{A}'\boldsymbol{A}\right)^{-1}\boldsymbol{A}'(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{\theta}^{\mathrm{im}}_{t|t-1}),\end{aligned}$$

which moves the parameter in a different direction compared with the ESD update. The ISD update also remains bounded as the learning rate η^{im} is increased.

Error bounds. For both filters in the above setup, Theorem 2 holds with

$$a^{\text{ex}} = (1 - \min\{\alpha\eta^{\text{ex}}, 2 - \beta\eta^{\text{ex}}\})^2 + \chi^2 L^2 \eta^{\text{ex}2}, \quad b^{\text{ex}} = (1 + 1/\chi^2)\sigma^2\eta^{\text{ex}}, \quad c^{\text{ex}} = 1, \quad d^{\text{ex}} = \sigma_\xi^2/\eta^{\text{ex}},$$
$$a^{\text{im}} = (1 + \alpha\eta^{\text{im}})^{-2}, \quad b^{\text{im}} = a^{\text{im}}\sigma^2\eta^{\text{im}}, \quad c^{\text{im}} = 1, \quad d^{\text{im}} = \sigma_\xi^2/\eta^{\text{im}}.$$

These values can be substituted into equations (18) and (19) to obtain finite-sample and asymptotic error bounds in the $\boldsymbol{P}^j$-norm, respectively. As we are interested in least squares, we multiply these bounds by $1/\lambda_{\min}(\boldsymbol{P}^j) = \eta^j$ to obtain (Euclidean) MSE bounds. Then we minimize these MSE bounds by choosing the learning rates η^j for $j \in \{\text{im}, \text{ex}\}$. For the ISD filter, we use the analytic minimizer $\eta_\star^{\text{im}} = 1/\rho_\star^{\text{im}}$ given in Corollary 2. For the ESD filter, we numerically minimize the MSE bound with respect to both the learning learning rate $\eta_\star^{\text{ex}}$ and Young's parameter $\chi = \chi_\star$.

Simulation setup. As in Cutler et al. (2023), our default parameters are $k = 50$, $n = 100$, $\alpha = \beta = 1$, $\sigma = 10$ and $\sigma_\xi = 1$. We simulate 1,000 paths of length $T = 500$. For each replication, we compute the final squared error $\|\boldsymbol{\theta}_{T|T}^j - \boldsymbol{\vartheta}_T\|^2$ for $j \in \{\text{im}, \text{ex}\}$. We compute MSEs by averaging over all replications.

Findings for ISD and ESD filters. Figure 1 shows the empirical MSEs of ISD and ESD filters for various values of t, η, σ_ξ, and β. The ISD filter consistently achieves lower MSEs than the ESD filter throughout the range of settings considered. The top-left plot shows that the ISD filter has the lowest empirical MSE and MSE bound for all time steps. The initial MSE for both filters is 100; this is because they are initialized in the origin, while the true process starts on a sphere of radius 10. The top-right plot shows the performance of both filters for different learning rates, illustrating that the advantage of ISD filters grows as the learning rate is increased. Vertical lines indicate the learning rates that minimize the MSE bounds, which in both cases lie slightly left of the learning rates that minimize the empirical MSEs. The bottom-left plot shows that the advantage of ISD filters grows when the state is more volatile (i.e., for larger values of σ_ξ). The bottom-right plot shows that the advantage of ISD filters also grows when observations are more informative (i.e., for larger values of β).

Three additional tracking algorithms. Next, we additionally consider three recent tracking algorithms: (i) the online version of Nesterov's cornerstone modern optimization method (ONM; Nesterov, 1983) as discussed in Nesterov (2018), (ii) the online gradient descent algorithm from Madden et al. (2021), and (iii) the stochastic gradient method from Cutler et al. (2023). For ONM, we follow the implementation provided in Madden et al. (2021, Section 6.1), where it showed strong empirical performance although no performance

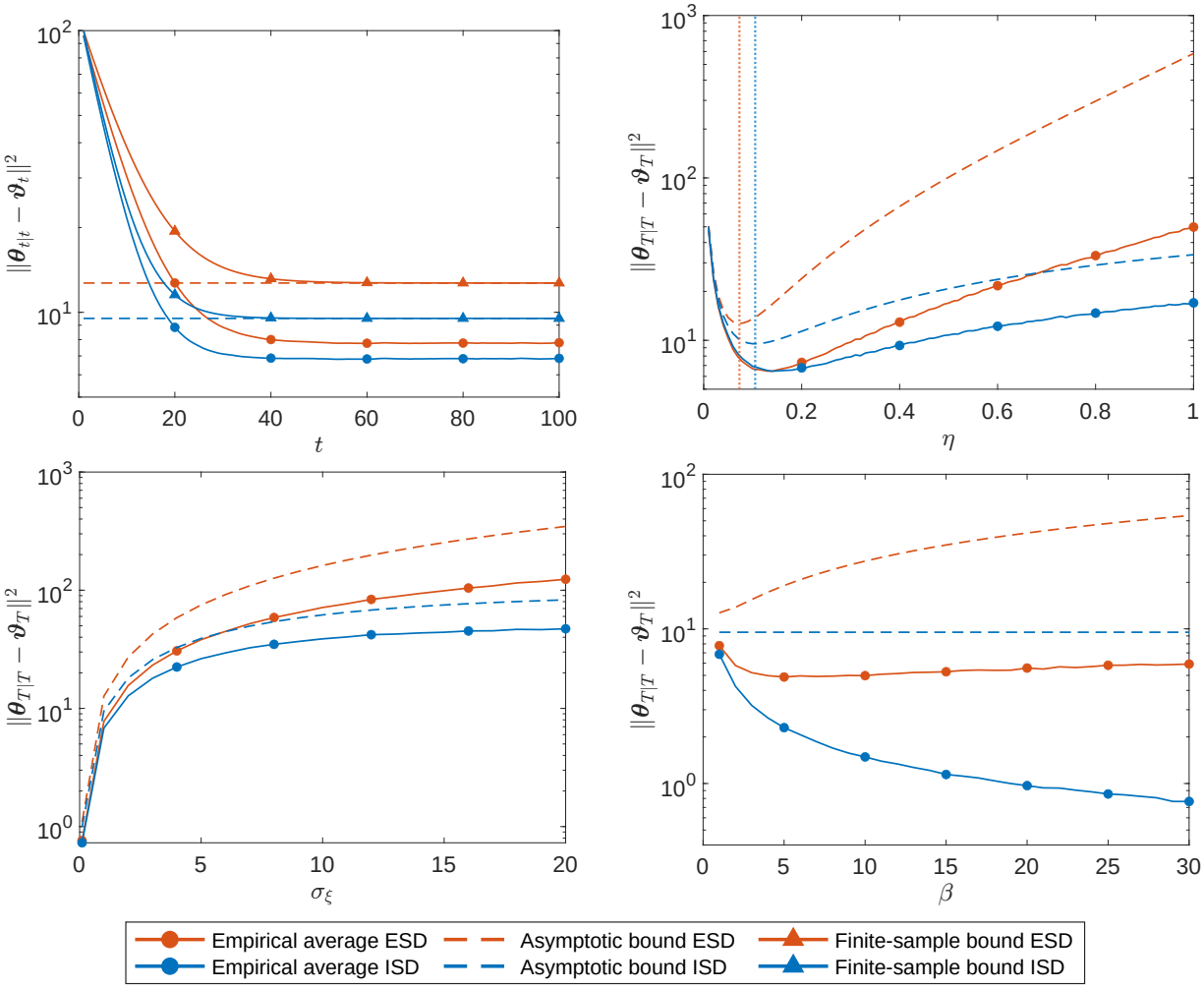

Figure 1: Semilog plots of empirical MSEs and MSE bounds (dotted) for least-squares recovery with respect to the time step t, learning rate η, state volatility σ_ξ, and Lipschitz gradient constant β, with the average errors computed at horizon $T = 500$ for the latter three plots. Empirical averages are computed over 1,000 replications. Unless stated otherwise, the parameters are $k = 50$, $n = 100$, $\alpha = \beta = 1$, $\sigma = 10$ and $\sigma_\xi = 1$, while learning rates η^j for $j \in \{\text{im}, \text{ex}\}$ are found by minimizing the asymptotic MSE bounds.

guarantees (in finite-dimensional spaces) are known. The algorithms by Madden et al. (2021) and Cutler et al. (2023) both use explicit gradient methods and, in this setting, are equivalent to our ESD filter—except, they employ different learning rates. These authors also derive error bounds and select their learning rates to minimize these bounds; for details, see Cutler et al. (2023, Theorem 15), and Madden et al. (2021, Theorem 3.1). Interestingly, Cutler et al. (2023) cap the learning rate at $1/(2\beta)$, while our learning-rate cap for the ESD filter is four times higher at $2/\beta$, which still guarantees stability via Theorem 1. When $\alpha = \beta$, ONM is identical to online gradient descent as in by Madden et al. (2021). The learning rate used in Madden et al. (2021) is $2/(\alpha + \beta)$, which differs from from ours and Cutler et al.'s in that it is independent of the gradient noise σ^2. The bounds in Madden

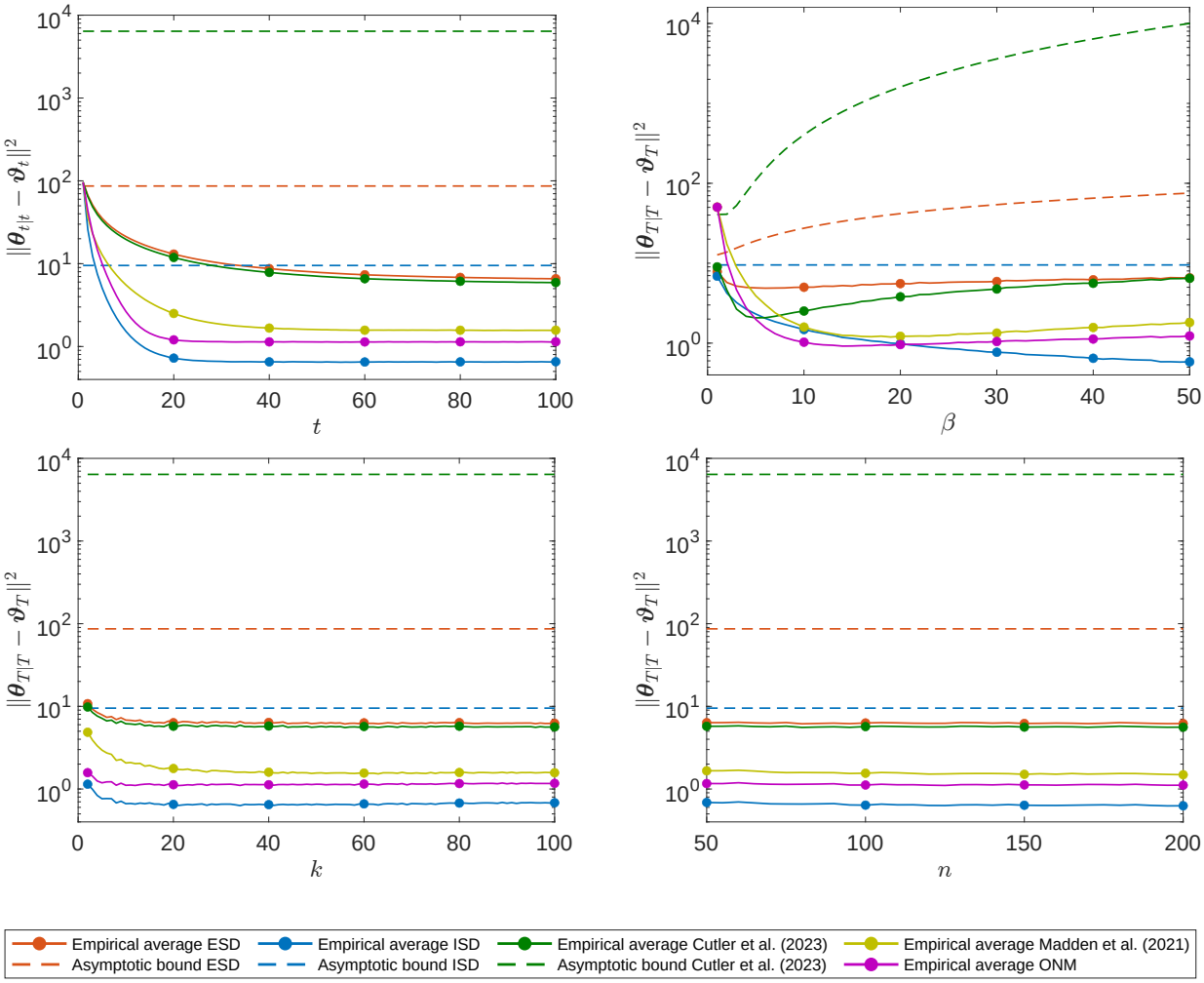

Figure 2: Semilog plots of guaranteed bounds and empirical tracking errors for least-squares recovery with respect to iteration t, Lipschitz gradient constant β, state dimension k, and observation dimension n, with the average errors computed at horizon $T = 500$ for the latter three plots. Empirical averages are computed over 1,000 trials. Default parameter values: $\alpha = 1, \beta = 40, \sigma = 10, \sigma_\xi = 1, \eta = \eta_\star$, $k = 50$, $n = 100$.

et al. (2021) use a different norm; hence, we do not show these in our graphs.

Comparison of five tracking algorithms. To investigate the differences between all five algorithms, we take a setting with a large disparity between α and β, taking $\alpha = 1$ and $\beta = 40$. For completeness, results for our default parameters (for which $\alpha = \beta = 1$ as in Cutler et al., 2023), are given in Appendix B.3. Figure 2 shows the MSEs of all five tracking algorithms and, when available, the asymptotic MSE bounds. We find that the ISD filter (i) outperforms all other methods across all time steps t, (ii) is the only algorithm that yields consistently lower MSEs when the Lipschitz constant β is increased, (iii) has the lowest MSE for all state dimensions k and observation dimensions n considered, (iv) provides the strongest (non-)asymptotic performance guarantees (i.e., lowest MSE bounds). Among the

alternatives, ONM performs best overall, aligning with findings in Madden et al. (2021), even though no guarantees for this algorithm are known.

5.2 Performance guarantees for 9 DGPs in Koopman et al. (2016)

DGPs. We take nine distributions from Koopman et al. (2016), listed in Table 1, which we combine with linear state dynamics $\vartheta_t = \phi_0 \vartheta_{t-1} + \xi_t$ initialized at $\vartheta_0 = 0$, where $\xi_t \sim \text{i.i.d.}(0, \sigma_\xi^2)$ is Student's t distributed with six degrees of freedom. We allow fat-tailed state increments, because our theoretical guarantees merely require the existence of two moments. For completeness, we also consider the Gaussian case (in Appendix B.4). The static parameters are $\phi_0 = 0.97$ and $\sigma_\xi \in \{0.15, 0.3, 0.6\}$ for a low-, medium-, and high-volatility setting. Koopman et al. (2016) investigated the low-volatility case (i.e., $\sigma_\xi = 0.15$) with Gaussian state innovations, finding (explicit) score-driven filters to be quite accurate. In this section we investigate not whether our MSE bounds are sharp (they are not); instead, we focus on the—as we will see, far-reaching—implications of their (un)boundedness.

Filters. We follow Koopman et al. (2016) in assuming the observation densities in Table 1 to be "correctly specified" in that our postulated density $p(\cdot|\theta_t)$ matches (the functional form of) the true density $p^0(\cdot|\vartheta_t)$; for exact specifications, see Appendix B.4. Some densities contain additional shape parameters (see Table B.2), which are assumed to be unknown and must be estimated by maximum likelihood (6). While ESD updates (2) are given in closed form (as the score is known in closed form), ISD updates (4) can be computed using standard numerical methods (for details, see Appendix B.6).

MSE bounds. For each density, Table 1 indicates the relevant values of α and β in Assumption 1, where $\alpha \geq 0$ indicates log concavity, while $\beta < \infty$ indicates a bounded Hessian. For the two non-concave cases, we impose $\rho > \alpha^-$ to ensure that Assumption 2 holds; in practice, this restriction is never binding. Assumption 3 holds automatically for all models as $\Theta = \mathbb{R}$. Based on the linear state equation, it can also be verified that Assumption 4 hold. Using the stated values of α and β, Table 1 indicates by check marks whether MSE bounds for the ISD and ESD filters exist (i.e., are finite), which is the case for nine and two models, respectively.

Simulation setting. For each DGP, we simulate 1,000 time series of length 10,000. The "in-sample" period of the first 1,000 observations is used for the estimation of static parameters (i.e., ω, ϕ, ρ and shape parameters), while the remaining "out-of-sample" period is used to compute MSEs of predictions $\{\theta^j_{t|t-1}\}$ for $j \in \{\text{im}, \text{ex}\}$ relative to true states $\{\vartheta_t\}$.

Findings. Table 1 shows that the ISD filter outperforms the ESD filter across all models and volatility settings. In the low-volatility setting (as in Koopman et al., 2016),

the differences are marginal. While the ESD filter diverged in the out-of-sample period for a single path of the Gaussian dependence model, for simplicity we ignore this path and report a finite MSE. (This divergence may explain why Koopman et al. (2016) further reduced the state volatility for both dependence models to $\sigma_\xi = 0.10$.) In the medium-volatility setting, the differences are substantial. For three models, the ESD filter is divergent in the out-of-sample period, with a substantial proportion of paths affected (e.g., $\sim$10% of paths for the Gaussian dependence model). We report these MSEs as being infinite. For two models with $\beta < \infty$, the differences remain marginal. In the high-volatility setting, the instability of the ESD filter is apparent for all models except those with $\beta < \infty$. Even when the ESD filter does not technically diverge (i.e., to infinity), some MSEs are very large indeed (e.g., $\sim 10^7$).

In sum, ESD filters remain competitive in all volatility scenarios only if an MSE bound is available. If this is not the case, ESD filters may diverge sooner or later; i.e., in the low-, medium- or high-volatility setting, depending on the model. These insights are apparently new in the filtering literature; e.g., the potential instability of (explicit) score-driven filters went unnoticed in Koopman et al. (2016), because they focused on the empirical performance (i.e., without offering theoretical guarantees) in a low-volatility setting. While our theory suggests that $\beta < \infty$ is often *sufficient* to obtain finite MSE bounds for the ESD filter, these simulation results suggest that $\beta < \infty$ is near *necessary*. Indeed, all our ESD-filter implementations with $\beta = \infty$ are unstable in some (sufficiently volatile) setting. While we focused on ESD filters with constant learning rates $\eta = \rho^{-1} > 0$ (which in the literature corresponds to "unit scaling"), these stability problems do not disappear (and may worsen) when using a time-varying learning rate; e.g., in the next subsection we show that these problems persist for the Poisson count model under all standard scaling options.

5.3 Poisson count model with various link and scaling functions

DGP. We consider count observations $y_t \in \mathbb{N}$ generated by a dynamic Poisson distribution $p^0(y_t|\mu_t) = \mu_t^{y_t} \exp(-\mu_t)/y_t!$ with mean $\mu_t = \exp(\vartheta_t)$, where positivity of $\mu_t > 0$ is ensured via an exponential link function. The latent process $\{\vartheta_t\}$ follows linear dynamics $\vartheta_t = 0.98\vartheta_{t-1} + \xi_t$ with Gaussian increments $\xi_t \sim$ i.i.d. $\mathrm{N}(0, \sigma_\xi^2)$. This model and slight variations have been extensively studied with notable applications including the UK van driver kills dataset (Harvey and Fernandes, 1989; Durbin and Koopman, 1997, 2000).

ESD filter. Explicit score-driven filters for the Poisson count model were considered in Gorgi (2018, sec. 5.2), Gorgi et al. (2024, sec. 4.3), and Davis et al. (2003, sec. 2.3). Following this literature, we assume that the Poisson distribution and the exponential

Table 1: Out-of-sample MSE of $\{\theta^j_{t|t-1}\}$ for $j \in \{\text{im}, \text{ex}\}$ relative to true states $\{\vartheta_t\}$.

DGP		**Assum. 1**		**MSE bound?**		**Low vol.** ($\sigma_\xi = 0.15$)		**Med. vol** ($\sigma_\xi = 0.30$)		**High vol.** ($\sigma_\xi = 0.60$)	
Type	Distribution	α	β	ISD	ESD	ISD	ESD	ISD	ESD	ISD	ESD
Count	Poisson	0	∞	✓	✗	.146	.149	.408	∞	1.744	∞
Count	Neg. Binomial	0	∞	✓	✗	.159	.160	.430	∞	1.681	∞
Intensity	Exponential	0	∞	✓	✗	.146	.150	.370	.899	1.064	$\sim 10^3$
Duration	Gamma	0	∞	✓	✗	.157	.162	.481	.577	1.324	$\sim 10^6$
Duration	Weibull	0	∞	✓	✗	.125	.128	.307	.345	0.797	$\sim 10^3$
Volatility	Gaussian	0	∞	✓	✗	.193	.199	.506	.647	1.513	$\sim 10^7$
Volatility	Student's t	0	$\frac{\nu+1}{8}$	✓	✓	.226	.226	.608	.615	1.559	1.612
Dependence	Gaussian	$-\frac{1}{4}$	∞	✓	✗	.237	.239†	.593	∞	1.581	∞
Dependence	Student's t	$-\frac{1}{4}$	$\frac{\nu+1}{4}$	✓	✓	.251	.251	.619	.624	1.515	1.548

Note: MSE = mean squared error. ISD = implicit score driven. ESD = explicit score driven. † For the Gaussian dependence model in the low-volatility setting, the ESD filter diverged in the out-of-sample period for single replication; for simplicity, we ignore this path and report a finite MSE.

link $\mu_t = \exp(\vartheta_t)$ are known to the researcher, so that they can be used for computing the score and its scaling. We consider all standard scaling options in the literature (i.e., $\zeta \in \{0, 1/2, 1\}$), including unit scaling and inverse (square root) Fisher scaling. However, since the scaled score $\exp(-\zeta\theta_{t|t-1})\,[y_t - \exp(\theta^{\text{ex}}_{t|t-1})]$ is not Lipschitz continuous in $\theta \in \mathbb{R}$ for any scaling, neither stability nor performance guarantees for the ESD filter are available.

ISD filter. For the ISD filter, we consider only a static learning rate (which corresponds to "unit scaling"). Although the Poisson distribution is again considered known, we investigate what happens if the relation between ϑ_t and μ_t is unknown. Specifically, we consider two cases, in which the researcher postulates (i) an exponential link $\mu^{\text{im}}_{t|t} = \exp(\theta^{\text{im}}_{t|t})$, which is correct, or (ii) a quadratic link $\mu^{\text{im}}_{t|t} = (\theta^{\text{im}}_{t|t})^2$, which is not. The parameter space Θ is $\mathbb{R}$ under the exponential link, or $\mathbb{R}_{\geq 0}$ under the quadratic link. In the latter case, simple parameter restrictions (i.e., positivity of ω and ϕ) ensure that the prediction step (3) maps $\mathbb{R}_{\geq 0}$ to itself. The ability to deal with bounded parameter spaces is unique to the ISD filter. Under the two link functions, the logarithmic Poisson distribution is either (i) concave in $\theta \in \mathbb{R}$ (i.e., $\alpha = 0$) or (ii) strongly concave in $\theta \in \mathbb{R}_{\geq 0}$ (with $\alpha = 2$). For both links, therefore, Assumption 1 holds. For the exponential link, the ISD update can be computed using Newton's optimization method (see Appendix B.6), while for the quadratic link a closed-form solution is available (see Appendix B.5). For the quadratic link, if the prediction lies in the interior of $\Theta = \mathbb{R}_{\geq 0}$, then the same holds for the update. Even if the

prediction and update both lie on the boundary (which is possible if $y_t = 0$), the first-order condition (1) is still satisfied; hence, Assumption 3 does not technically hold, but can be circumvented. In both cases Assumption 4 holds with $\sigma^2 = 4$ (for details, Appendix B.5).

In sum, performance guarantees for both versions of the ISD filter are available. When using the correct (exponential) link function, these bounds concern the tracking error relative to the true states, $\{\vartheta_t\}$. However, when using the incorrect (quadratic) link, these bounds relate to the tracking error relative to the pseudo-true states $\{\theta_t^\star\} = \{\exp(\vartheta_t/2)\}$. The bound in the latter case involves one free parameter that we choose as in Corollary 1 to yield the lowest permissible bound. In this case, the minimized MSE bound further depends on $q^2 = \sup_t \mathbb{E}[\|\theta_t^\star - \theta_{t-1}^\star\|^2] < \infty$ and $s_\omega^2 = \sup_t \mathbb{E}[\|\theta_t^\star - \omega\|] < \infty$. For computing the bound, these quantities are assumed to be known; this is common in the stochastic-optimization literature (e.g., Nesterov, 2018, p. 8).

Simulation setup. We vary the state variation σ_ξ in the range $(0, 0.5]$. For each value σ_ξ, we simulate 500 time series of length 10,000. The "in-sample" period consisting of the first 1,000 observations is available for the estimation of the static parameters, while the remaining "out-of-sample" period is used to asses the filtering performance.

Parameter estimation. Filters with exponential link functions are implemented with Assumption 5; i.e., the true parameters $\omega_0 = 0$ and $\phi_0 = 0.98$ are used in the prediction step (3) (i.e., we set $\omega^j = \omega_0$ and $\phi^j = \phi_0$ for $j \in \{\text{im}, \text{ex}\}$), while the learning rates η^j for $j \in \{\text{im}, \text{ex}\}$ are estimated using (6). For the ISD filter with the quadratic link, all static parameters in the filter (i.e., ω^{im}, ϕ^{im}, η^{im}) are estimated using (6).

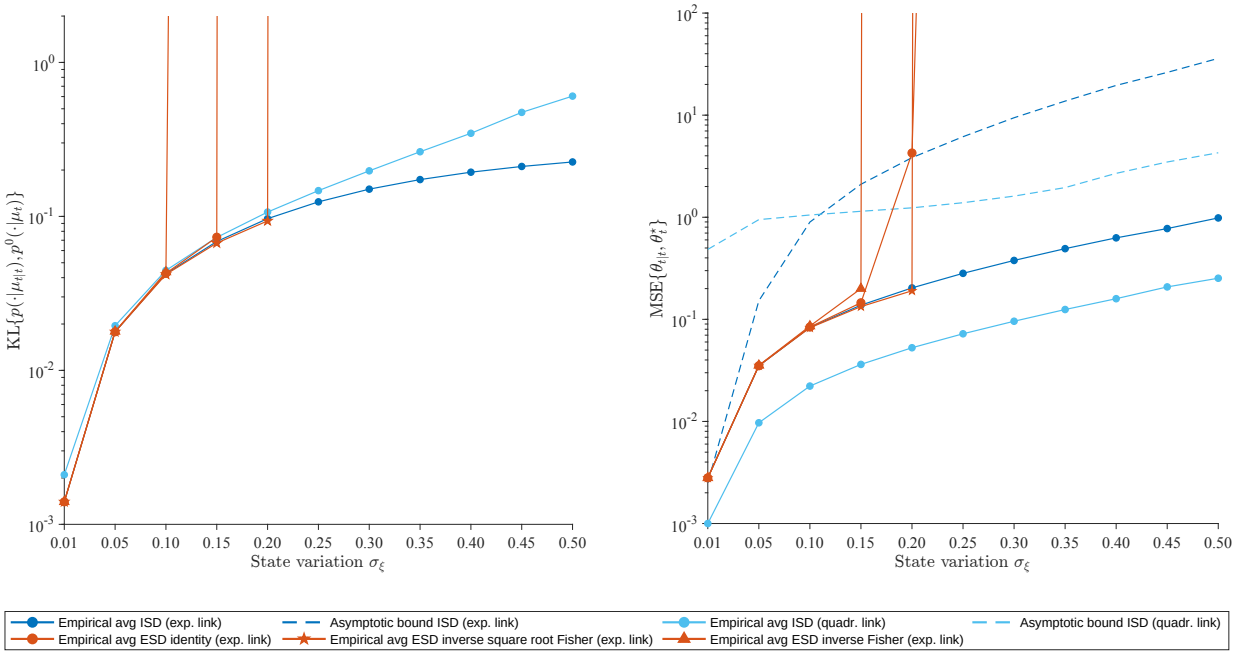

Figure 3: Plots of the out-of-sample empirical errors and guaranteed bound for tracking (the logarithm of the rate of) a dynamic Poisson distribution, i.e., the true state $\{\vartheta_t\}$, with respect to its variation σ_ξ. The left plot shows the Kullback-Leibler (KL) divergence between the postulated and true densities $p(\cdot|\mu_{t|t})$ and $p^0(\cdot|\mu_t)$, where $\mu_{t|t}$ and μ_t denote the filtered and true rates, respectively. The right plot shows the MSE between the filtered state, $\theta_{t|t}$ and the pseudo-true state $\theta_t^\star$ ($= \vartheta_t$ if Assumption 5 holds). Empirical averages are computed over 500 trials. Default parameter values: $\omega_0 = 0$, $\phi_0 = 0.98$, $\alpha = 2, \beta = \infty$.

Findings. The left panel of Figure 3 presents the out-of-sample Kullback-Leibler (KL) divergence between the filtered distribution $p(\cdot|\mu_{t|t})$ and the true distribution $p^0(\cdot|\mu_t)$, expressed as $\mathrm{KL}\{p(\cdot|\mu_{t|t}), p^0(\cdot|\mu_t)\} = \mu_t \log(\mu_t/\mu_{t|t}) + \mu_{t|t} - \mu_t$. Because for each filter the postulated distribution is Poisson, the divergence depends only on how the filtered mean $\mu_{t|t}^j$ for $j \in \{\mathrm{im}, \mathrm{ex}\}$ compares with the true mean μ_t. For low state volatilities, all filters achieve similar KL divergences. For sufficiently large state variations, however, all three ESD filters are unstable. The ISD filters remain stable for all state variations, even as their estimated learning rates are considerably larger (for estimation details, see Appendix B.5).

The right-hand panel of Figure 3 shows the out-of-sample MSEs of filtered states $\{\theta_{t|t}^j\}$, relative to true states $\{\vartheta_t\}$ (in the case of an exponential link) or pseudo-true states $\{\theta_t^\star = \exp(\vartheta_t/2)\}$ (in the case of a quadratic link). Although they are shown in the same plot, we can only meaningfully compare MSE values associated with the same link function. As the state variability increases, all three ESD filters become unstable. In line with our theoretical guarantees, both ISD filters remain stable for all state variations.

6 Empirical illustration: T-bill rate spreads

This section illustrates empirical advantages of ISD filters even under boundedness of the Hessian. We focus on location models as in Example 2, in which the observations are contaminated with Student's t distributed observation noise. While the Student's t violates log-concavity, the Hessian is bounded. The associated ESD filter, introduced in Harvey and Luati (2014), has found widespread application; see Caivano et al. (2016, sec. 3.1), Blasques et al. (2018, sec. 6.3), Blasques et al. (2022, sec. 5), Artemova et al. (2022a) and Gorgi et al. (2024, sec. 4.5). Here we formulate an ISD version of the filter and demonstrate that it is associated with a more favorable impact curve, which avoids the "zig-zagging" behavior of the filtered path when the data are constant, while enhancing the filter's responsiveness when the data are variable. We illustrate these advantages in an empirical illustration based on Artemova et al. (2022a) involving U.S. treasury-bill (T-bill) rate spreads. We show that these advantages further extend to another model used in Artemova et al. (2022a), formulated by Caivano and Harvey (2014), where the observation noise follows an exponential generalized beta distribution of the second kind (EGB2).

6.1 Student's t location model: ESD and ISD versions

The ESD update for the Student's t location model (e.g., Harvey and Luati, 2014, eq. 5) is

$$\theta^{\text{ex}}_{t|t} = \theta^{\text{ex}}_{t|t-1} + \eta\,(y_t - \theta^{\text{ex}}_{t|t-1})\left[1 + \frac{1}{\nu}\left(\frac{y_t - \theta^{\text{ex}}_{t|t-1}}{\varsigma}\right)^2\right]^{-1} = (1 - w^{\text{ex}}_{t|t})\theta^{\text{ex}}_{t|t-1} + w^{\text{ex}}_{t|t} y_t, \qquad (26)$$

with learning rate $\eta > 0$, scale parameter $\varsigma > 0$, and degrees of freedom $\nu > 0$. The second equality demonstrates that $\theta^{\text{ex}}_{t|t}$ can be written as a weighted average of the prediction and the observation y_t, where the weight associated with y_t is $w^{\text{ex}}_{t|t} := \eta[1 + \nu^{-1}(y_t - \theta^{\text{ex}}_{t|t-1})^2/\varsigma^2]^{-1} \geq 0$. Here, $w^{\text{ex}}_{t|t} \leq \eta$ with equality in the limit $\nu \to \infty$. The update is complemented with the usual prediction step, i.e. $\theta^{\text{ex}}_{t+1|t} = \omega(1-\phi) + \phi\theta^{\text{ex}}_{t|t}$, where $|\phi| \leq 1$. The combination of both steps yields the standard prediction-to-prediction recursion in ESD models, with the only difference that we have the parameter combination $\eta\phi$ in front of the score; up to re-parametrization of η, this is equivalent.

Unfortunately, the ESD update (26) does not guarantee that $\theta^{\text{ex}}_{t|t}$ lies in the interval between the prediction $\theta^{\text{ex}}_{t|t-1}$ and the observation y_t. Unless the learning rate η is restricted to be less than unity, the weight $w^{\text{ex}}_{t|t}$ may exceed one, which means that we "go short" in the prediction and "overshoot" the observation. Mechanically, this may not be a problem; e.g. Harvey and Luati (2014, p. 1117) state that their learning rate (which is $\eta\phi$ in our notation) can be estimated "without unity imposed as an upper bound". In an empirical application

involving treasury bill spreads, Artemova et al. (2022a, Table 1) estimate learning rates considerably above one (e.g. $\eta \approx 2.2$ and $\eta\phi \approx 1.6$). However, as is visible in Figure 3 of Artemova et al. (2022a), and also in our Figure 4 below, this means that the ESD filter frequently overshoots the data, leading to "zig-zagging" behavior of the filtered path, in particular when the data are stable. The same problem is apparent for the location model based on the EGB2 distribution (Artemova et al., 2022a, Fig. 3). For the Student's t model, imposing the constraint $\eta \leq 1$ would seem to present an obvious solution; however, the model fit would be reduced and the filter's responsiveness suppressed.

Here we present an alternative solution to the overshooting problem, which is to formulate an ISD version of the model. The ISD update automatically ensures that, for any learning rate $\eta > 0$, $\theta^{\text{im}}_{t|t}$ lies in the interval between $\min\{\theta^{\text{im}}_{t|t-1}, y_t\}$ and $\max\{\theta^{\text{im}}_{t|t-1}, y_t\}$. To see why, we note that the Student's t log-density $\log p(y_t|\theta)$ in optimization (4) is maximized at $\theta = y_t$, while the penalty is increasing in the distance $|\theta - \theta_{t|t-1}|$; hence, updating beyond y_t cannot be optimal. The implicit version of the ESD update (26) reads

$$\theta^{\text{im}}_{t|t} \;=\; \theta^{\text{im}}_{t|t-1} \;+\; \eta\,(y_t - \theta^{\text{im}}_{t|t}) \left[1 + \frac{1}{\nu}\left(\frac{y_t - \theta^{\text{im}}_{t|t}}{\varsigma}\right)^2\right]^{-1}, \tag{27}$$

where the implicit update $\theta^{\text{im}}_{t|t}$ appears on both sides of the equation. The implicit update (27) can be analyzed by substituting $\theta^{\text{im}}_{t|t} = (1 - w^{\text{im}}_{t|t})\theta^{\text{im}}_{t|t-1} + w^{\text{im}}_{t|t} y_t$ for some weight $w^{\text{im}}_{t|t} \in (0, 1)$ to be found. After some algebra, making this substitution yields

$$\frac{1}{\nu}\left(\frac{y_t - \theta^{\text{im}}_{t|t-1}}{\varsigma}\right)^2 (1 - w^{\text{im}}_{t|t})^2\, w^{\text{im}}_{t|t} \;+\; (\eta + 1)\, w^{\text{im}}_{t|t} \;-\; \eta \;=\; 0, \tag{28}$$

which is a cubic equation for $w^{\text{im}}_{t|t}$. At most three distinct solutions (of which at least one is real valued) exist, which are available in standard software packages. When $\nu \to \infty$ or $y_t - \theta^{\text{im}}_{t|t-1} \to 0$, in which case the quadratic and cubic terms disappear, we get $w^{\text{im}}_{t|t} \to \eta/(1 + \eta)$. In fact, $w^{\text{im}}_{t|t} \leq \eta/(1 + \eta) < 1$ for any $\nu < \infty$, such that the update for the Student's t is always more conservative than in the Gaussian case. The weight being bounded above by $\eta/(1 + \eta)$ also implies that the update cannot surpass y_t. These desirable properties are not shared by the ESD update (26). The reason for the possible multiplicity of solutions to (28) is that Student's t distribution fails to be log concave; recall from Example 2 that $\alpha = -1/8$. In theory, up to three stationary points could exist, corresponding to one local minimum and two local maxima. If multiple stationary points exist, the associated function values may be compared to yield the global maximum. In the unlikely event that both local maxima yield identical values, a tie breaker can be used.

Table 2: Estimated parameters and log-likelihood values for T-bill spread data

Distribution	Filter	ω	ϕ	η	ς^2	$\nu, \varkappa$	LL	Thrm. 1
Student's t	ESD	1.234	0.714	2.194	0.516	2.632	-370.6	✓
	ISD	0.944	0.751	23.713	0.387	2.061	-364.4	✗
EGB2	ESD	1.114	0.714	0.441	1.401	0.173	-378.8	✗
	ISD	0.970	0.722	2.351	1.336	0.016	-369.9	✓

Note: Parameter estimates and log-likelihood (LL) values for location models with errors following a Student's t distribution (with parameter $\nu > 0$) or symmetric exponential generalized beta distribution of the second kind (EGB2, with parameter $\varkappa > 0$) for implicit (ISD) and explicit score-driven (ESD) filters. Tick marks indicate whether the sufficient conditions for invertibility in Theorem 1 are satisfied.

The multiplicity of solutions is ruled out if Assumption 2 holds, which imposes $\rho > 1/8$ or $\eta < 8$. In this case, equation (28) has a single real-valued solution, which lies in $(0, 1)$.

6.2 Parameter estimates and invertibility

Following Artemova et al. (2022a), we consider differences between 3-month and 6-month T-bill rates from the FRED-QD database (series "TB6M3M") scaled by a factor of 10, consisting of 249 quarterly observations from 1959-Q1 to 2021-Q1. Table 2 shows the log-likelihood values and parameter estimates for ESD and ISD level models using the Student's t distribution (with shape parameter $\nu > 0$) and the symmetric EGB2 distribution (with shape parameter $\varkappa > 0$). For both distributions, the implicit version of the filter achieves higher likelihood values. Moreover, in both settings, the estimated learning rates for the implicit filter considerably exceed those of the explicit version. For the ISD filter with a Student's t distribution, the learning rate is estimated at $\hat{\eta} \approx 24 > 8$, such that multiple real-valued roots of equation (28) may exist, depending on y_t. This possibility materialized only when the filter encountered the extreme value of 10.6 on 1 September 1982 (see Figure 4 below), when the algorithm selected the stationary point associated with the highest value. Learning rates for the ESD filter are estimated to be considerably lower at $\hat{\eta} \approx 2.2$ for the the Student's t distribution and $\hat{\eta} \approx 0.44$ for the EGB2. However, these estimates still violate the restrictions that would prevent overshooting, which read $\hat{\eta} \leq 1$ for the Student's t and $\hat{\eta} \leq 1/(\hat{\varkappa}\psi'(\hat{\varkappa})) \approx 0.166$ for the EGB2, where $\psi'(\cdot)$ is the trigamma function.

The last column of Table 2 indicates, for each calibrated model, whether the relevant sufficient condition for invertibility in Theorem 1 is satisfied. For the Student's t location model as in Example 2, Assumption 1 holds with $\alpha = -1/8$ and $\beta = 1$. Sufficient condition (13) for the ESD filter then reads $\hat{\eta} < \min\{8, 3.20, 2.40\}$, which is satisfied as $\hat{\eta} = 2.19$.

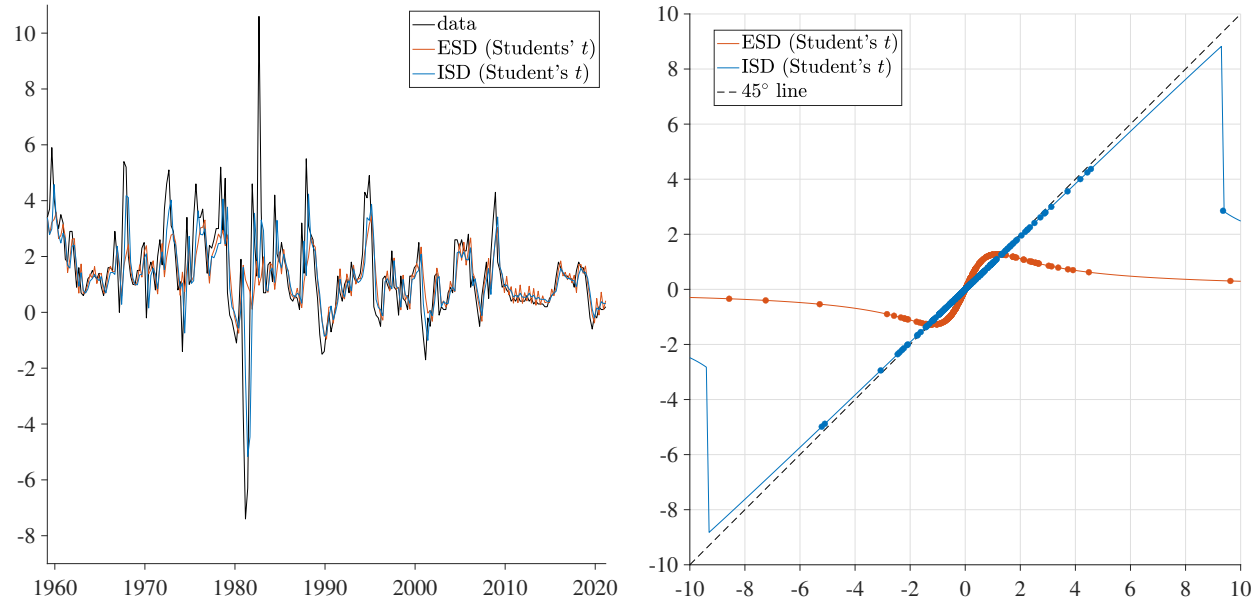

Figure 4: Time-series data of T-bill rate spreads modeled using a Student's t distribution with ESD and ISD predictions $\{\theta^j_{t|t-1}\}_t$ for $j \in \{\text{ex}, \text{im}\}$ (left) and associated impact curves (right), which shows $y_t - \theta^j_{t|t-1}$ (horizontally) versus $\theta^j_{t|t} - \theta^j_{t|t-1}$ (vertically) for $j \in \{\text{ex}, \text{im}\}$.

Sufficient condition (14) for the ISD filter, on the other hand, includes the requirement $\hat{\eta} < 8$, which is clearly violated. Interestingly, the data suggests that Assumption 2 should *not* hold: the ISD filter fits the data best when optimization (4) fails to be strongly concave. It seems this non-concavity enables the ISD filter to ignore one particular outlier (again, the value of 10.6 on 1 September 1982), which may be useful in practice even if it complicates the theory; for further discussion, see below.

For the logarithmic (symmetric) EGB2 density multiplied by ς^2 (e.g. Artemova et al., 2022a, eq. 6), Assumption 1 holds with $\alpha = 0$ (the density is log-concave) and $\beta = \varkappa\psi'(\varkappa)$ (the Hessian is bounded), with associated estimate $\hat{\beta} = \hat{\varkappa}\psi'(\hat{\varkappa}) \approx 6.016$ (using $\hat{\varkappa} \approx 0.173$). Sufficient condition (12) for the ESD filter becomes $\hat{\phi}^2(\hat{\beta}\hat{\eta} - 1)^2 < 1$, which fails to hold as the left-hand side equals 1.395. Sufficient condition (11) for invertibiltiy of the ISD filter with the EGB2 distribution, on the other hand, is trivially satisfied as $\alpha = 0$ and $\hat{\phi} < 1$.

6.3 Time-series behavior and impact curves

To investigate the time-series behavior resulting from the estimates in Table 2, the left-hand panel of Figure 4 shows predicted location parameters $\{\theta^j_{t|t-1}\}$ for $j \in \{\text{ex}, \text{im}\}$ based on the Student's t distribution. The right-hand panel displays the associated impact curves, plotting prediction errors $y_t - \theta^j_{t|t-1}$ (horizontally) against the "impact" $\theta^j_{t|t} - \theta^j_{t|t-1}$ (vertically) for $j \in \{\text{ex}, \text{im}\}$. The ESD predicted path is relatively unresponsive during volatile

times such as the period $1960-90$s, almost entirely ignoring several upward and downward spikes, while being overly responsive from 2010 onward, when the ESD filter "zig-zags" around the data. This behavior is due to the specific shape of the impact curve, which is roughly linear around the origin with a slope of $\hat{\eta} \approx 2.2 > 1$ (compare with the 45° line). For absolute prediction errors smaller than ∼1.3 (as happens for ∼80% of the observations), the filtered parameter $\theta^{\text{ex}}_{t|t}$ "overshoots" the observation y_t in that the parameter adjustment exceeds the prediction error, i.e. $|\theta^{\text{ex}}_{t|t} - \theta^{\text{ex}}_{t|t-1}| > |y_t - \theta^{\text{ex}}_{t|t-1}|$. On the other hand, the re-descending nature of the impact curve means that observations deviating more than ∼1.3 from predictions are largely ignored.

The ISD filter, in contrast, is associated with a roughly linear impact curve for absolute prediction errors all the way up to ∼9 (rather than up to ∼1 for the ESD filter). Due to this increased responsiveness, the ISD filter is better able to track peaks and troughs during the volatile period 1960-90. Even as the learning rate is estimated at $\hat{\eta} \approx 24$, the filter steadily tracks the data during the calm period starting around 2010; this stability is due to the slope of the impact curve around the origin being $\hat{\eta}/(1+\hat{\eta}) < 1$, ensuring that the update never surpasses y_t. The estimated impact curve suggests that a single outlier is to be largely ignored: the value of 10.6 on 1 September 1982, also mentioned above, which is associated with a prediction error of ∼9.5 for both filters (see the right-most dot on both impact curves). The discontinuity in the impact curve of the ISD filter is due to the emergence of three real-valued roots of equation (28), with the best one being selected. Whether the impact curve should be discontinuous (and, if so, where this discontinuity should be) was endogenously determined by estimating the model. Both filters also handle sizable negative prediction errors differently, with the ISD filter again being more responsive.

In all, the (in-sample) mean squared error improvement of ISD over ESD predictions is ∼18%, which may be economically significant. The results for the EGB2 distribution are qualitatively similar; except the news impact curves for both filters are (continuous and) monotone increasing rather than redescending (for details, see Appendix B.7).

7 Conclusion

This article established theoretical guarantees for the stability and filtering performance of multivariate explicit and implicit score-driven (ESD and ISD) filters under potential model misspecification. First, we presented novel sufficient conditions for exponential stability (i.e., invertibility) of the multivariate filtered parameter path. These conditions do not require log-concavity of the postulated observation density and are agnostic with regard

to the DGP, making them easily verifiable in practice. The conditions needed to prove stability for ESD filters were shown to be more stringent, requiring the score to be Lipschitz continuous in the parameter. Second, we combined these sufficient conditions for filter stability with mild conditions on the DGP to yield guarantees for both filters' tracking performance. Specifically, we provided analytic (non-)asymptotic mean squared error (MSE) bounds that relate the filtered (or predicted) parameter path to the pseudo-true parameter path. The asymptotic bounds can be minimized, often analytically, with respect to parameters of interest (e.g., the learning rate). In some special cases, these minimized bounds are exact and thus cannot be further improved.

In extensive Monte Carlo studies, we validated these theoretical findings and demonstrated the advantages of ISD over ESD filters. In a high-dimensional linear model, our new filtering bounds improve existing bounds by several orders of magnitude. In a wide range of non-linear settings, we showed that ISD filters are stable even when, due to non-Lipschitz-continuous scores, ESD filters are not. We also demonstrated that, when no (finite) MSE bounds are available and the underlying state is sufficiently volatile, ESD filters may be divergent. This finding is consistent with our theory development, but appears to be novel in the time-series literature. In an empirical illustration involving U.S. Treasury-bill rates, we demonstrated that a new implicit version of a popular score-driven location model with fat-tailed observation noise considerably improves tracking accuracy.

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*. Oxford University Press.

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*. Oxford University Press.

Bernardi, E., A. Lanconelli, C. S. Lauria, and B. T. Perçin (2024). Non trivial optimal sampling rate for estimating a Lipschitz-continuous function in presence of mean-reverting Ornstein-Uhlenbeck noise. https://arxiv.org/pdf/2405.10795.

Blasques, F., P. Gorgi, S. J. Koopman, and O. Wintenberger (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics 12*, 1019–1052.

Blasques, F., J. van Brummelen, S. J. Koopman, and A. Lucas (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics 227*(2), 325–346.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). Optimization methods for large-scale machine learning. *SIAM Review 60*(2), 223–311.

Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization 31*(4), 942–959.

Brownlees, C. and J. Llorens-Terrazas (2024). Empirical risk minimization for time series: Nonparametric performance bounds for prediction. *Journal of Econometrics 244*(1), 105849.

Caivano, M. and A. Harvey (2014). Time-series models with an EGB2 conditional distribution. *Journal of Time Series Analysis 35*(6), 558–571.

Caivano, M., A. Harvey, and A. Luati (2016). Robust time series models with trend and seasonal components. *SERIEs 7*, 99–120.

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.

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

Cutler, J., D. Drusvyatskiy, and Z. Harchaoui (2023). Stochastic optimization under distributional drift. *Journal of Machine Learning Research 24*(147), 1–56.

Davis, R. A., W. T. Dunsmuir, and S. B. Streett (2003). Observation-driven models for Poisson counts. *Biometrika 90*(4), 777–790.

Duchi, J. C. (2018). Introductory lectures on stochastic optimization. *The Mathematics of Data 25*, 99–186.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika 84*(3), 669–684.

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society Series B: Statistical Methodology 62*(1), 3–56.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*, Volume 38. Oxford University Press.

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. (2018). Integer-valued autoregressive models with survival probability driven by a stochastic recurrence equation. *Journal of Time Series Analysis 39*(2), 150–171.

Gorgi, P., C. Lauria, and A. Luati (2024). On the optimality of score-driven models. *Biometrika 111*(3), 865–880.

Guo, L. and L. Ljung (1995). Exponential stability of general tracking algorithms. *IEEE Transactions on Automatic Control 40*(8), 1376–1387.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*, Volume 52. Cambridge University Press.

Harvey, A. C. (2022). Score-driven time series models. *Annual Review of Statistics and Its Application 9*(1), 321–342.

Harvey, A. C. and C. Fernandes (1989). Time series models for count or qualitative observations. *Journal of Business & Economic Statistics 7*(4), 407–417.

Harvey, A. C. and A. Luati (2014). Filtering with heavy tails. *Journal of the American Statistical Association 109*(507), 1112–1122.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM Review 23*(1), 53–60.

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis*. Cambridge University Press.

Jungers, R. (2009). *The Joint Spectral Radius: Theory and Applications*, Volume 385. Springer.

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.

Lambert, M., S. Bonnabel, and F. Bach (2022). The recursive variational Gaussian approximation (R-VGA). *Statistics and Computing 32*(10), 1–24.

Lanconelli, A. and C. S. Lauria (2024). Maximum likelihood with a time varying parameter. *Statistical Papers 65*(4), 2555–2566.

Lange, R.-J. (2024a). Bellman filtering and smoothing for state–space models. *Journal of Econometrics 238*(2), 105632.

Lange, R.-J. (2024b). Short and simple introduction to Bellman filtering and smoothing. *arXiv preprint arXiv:2405.12668*.

Lange, R.-J., B. van Os, and D. J. van Dijk (2024). Implicit score-driven filters for time-varying parameter models. https://ssrn.com/abstract=4227958.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*. Springer.

Liu, D. C. and J. Nocedal (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming 45*(1), 503–528.

Madden, L., S. Becker, and E. Dall'Anese (2021). Bounds for the tracking error of first-order online optimization methods. *Journal of Optimization Theory and Applications 189*, 437–457.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization 19*(4), 1574–1609.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR 269*, 543–547.

Nesterov, Y. (2003). *Introductory Lectures on Convex Optimization: A Basic Course*, Volume 87. Springer.

Nesterov, Y. (2018). *Lectures on Convex Optimization*, Volume 137. Springer.

Ollivier, Y. (2018). Online natural gradient as a Kalman filter. *Electronic Journal of Statistics 12*, 2930–2961.

Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics 21*(1), 124–127.

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.

Straumann, D. and T. Mikosch (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Annals of Statistics 34*, 2449–2495.

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

Wilson, C., V. V. Veeravalli, and A. Nedić (2019). Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control 64*(2), 496–509.

Appendix to "Stability and performance guarantees for misspecified multivariate score-driven filters"

A Proofs

Preliminaries

Throughout our proofs, we make extensive use of the squared weighted Euclidean norm $\|\boldsymbol{x}\|^2_{\boldsymbol{W}} := \boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}$ for $\boldsymbol{x} \in \mathbb{R}^k$ and positive definite $\boldsymbol{W} \in \mathbb{R}^{k\times k}$. With slight abuse of notation, we also use the shorthand $\|\boldsymbol{x}\|^2_{\boldsymbol{W}}$ for $\boldsymbol{W}$ that are not positive definite; in this case, the expression remains well defined but strictly speaking does not constitute a norm.

The notation $\boldsymbol{A}^{1/2}$ for positive semidefinite $\boldsymbol{A} \in \mathbb{R}^{k\times k}$ refers to the symmetric matrix square root that can be obtained via eigendecomposition. Specifically, let $\boldsymbol{V} \in \mathbb{R}^{k\times k}$ denote the matrix of eigenvectors of $\boldsymbol{A}$ and let $\boldsymbol{D} \in \mathbb{R}^{k\times k}$ denote the diagonal matrix of eigenvalues in corresponding order such that $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'$; then, $\boldsymbol{A}^{1/2} := \boldsymbol{V}\tilde{\boldsymbol{D}}\boldsymbol{V}'$, where $\tilde{\boldsymbol{D}} \in \mathbb{R}^{k\times k}$ is a diagonal matrix containing the square root eigenvalues such that $\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}} = \boldsymbol{D}$. From this definition and the orthogonality of $\boldsymbol{V}$ it is immediately apparent that $\boldsymbol{A}^{1/2}$ is symmetric and that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{V}\tilde{\boldsymbol{D}}\boldsymbol{V}'\boldsymbol{V}\tilde{\boldsymbol{D}}\boldsymbol{V}' = \boldsymbol{V}\tilde{\boldsymbol{D}}\boldsymbol{I}_k\tilde{\boldsymbol{D}}\boldsymbol{V}' = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}' = \boldsymbol{A}$. It can also be established that $\boldsymbol{A}^{1/2}$ is the unique symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$, see Horn and Johnson (2012, Thm. 7.2.6a).

A.1 Example 1

Linear Gaussian observation equation. Observations $\boldsymbol{y}_t \in \mathbb{R}^n$ are generated by

$$\boldsymbol{y}_t = \boldsymbol{d} + \boldsymbol{Z}\,\boldsymbol{\vartheta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \text{i.i.d.N}(\boldsymbol{0}_n, \boldsymbol{\Sigma}_\varepsilon), \tag{A.1}$$

where $\boldsymbol{d}, \boldsymbol{\varepsilon}_t \in \mathbb{R}^n$, $\boldsymbol{Z} \in \mathbb{R}^{n\times k}$, $\boldsymbol{\vartheta}_t \in \mathbb{R}^k$ and $\boldsymbol{\Sigma}_\varepsilon \succ \boldsymbol{O}_n$. If the postulated density $p(\cdot|\boldsymbol{\theta})$ matches the true density $p^0(\cdot|\boldsymbol{\vartheta})$, then the log-likelihood function and score are

$$\begin{aligned} \text{log density:} \qquad & \ell(\boldsymbol{y}|\boldsymbol{\theta}) \propto -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{d}-\boldsymbol{Z}\boldsymbol{\theta})'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}-\boldsymbol{d}-\boldsymbol{Z}\boldsymbol{\theta}), \\ \text{score:} \qquad & \nabla\ell(\boldsymbol{y}|\boldsymbol{\theta}) = \boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}-\boldsymbol{d}-\boldsymbol{Z}\boldsymbol{\theta}). \end{aligned} \tag{A.2}$$

Let the predicted parameter $\boldsymbol{\theta}_{t|t-1} \in \mathbb{R}^k$ be given and fixed, which is used to compute both the ISD update (1) and the ESD update (2); hence, we omit the superscript on $\boldsymbol{\theta}_{t|t-1}$.

Kalman's covariance update. Let Kalman's predicted and updated covariance matrices $\boldsymbol{P}^{\text{KF}}_{t|t-1} \succ \boldsymbol{O}_k$ and $\boldsymbol{P}^{\text{KF}}_{t|t} \succ \boldsymbol{O}_k$ be given and fixed. They are related via

$$\begin{aligned} \boldsymbol{P}^{\text{KF}}_{t|t} &= \boldsymbol{P}^{\text{KF}}_{t|t-1} - \boldsymbol{P}^{\text{KF}}_{t|t-1}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{P}^{\text{KF}}_{t|t-1}\boldsymbol{Z}' + \boldsymbol{\Sigma}_\varepsilon)^{-1}\boldsymbol{Z}\boldsymbol{P}^{\text{KF}}_{t|t-1}, \\ &= [(\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1} + \boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z}]^{-1}. \end{aligned} \tag{A.3}$$

The first line gives the standard covariance-matrix updating step; see Harvey (1989, p. 106) or Durbin and Koopman (2012, p. 86). The second line is used below; it follows from the Woodbury matrix-inversion lemma (see equation (A.8) below) and can be found in several places (e.g. Fahrmeir, 1992, p. 504, Lambert et al., 2022, eq. 36).

Kalman update as ISD update. We take the ISD update (1) with $\boldsymbol{H}^{\text{im}}_t = \boldsymbol{P}^{\text{KF}}_{t|t-1}$ and substitute the score (A.2) to yield

$$\begin{aligned} \boldsymbol{\theta}^{\text{im}}_{t|t} &= \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t-1}\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}^{\text{im}}_{t|t}), \\ &= \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t-1}\,\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}^{\text{im}}_{t|t}). \end{aligned} \tag{A.4}$$

Pre-multiplying by $(\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1}$ and isolating $\boldsymbol{\theta}^{\text{im}}_{t|t}$ on the left-hand side, we obtain

$$\begin{aligned}
\boldsymbol{\theta}^{\text{im}}_{t|t} &= (\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z} + (\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1})^{-1}[(\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1}\boldsymbol{\theta}_{t|t-1} + \boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}_t - \boldsymbol{d})], \\
&= (\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z} + (\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1})^{-1}[\{(\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1} + \cancel{\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z}} - \cancel{\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z}}\}\boldsymbol{\theta}_{t|t-1} + \boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}_t - \boldsymbol{d})], \\
&= \boldsymbol{\theta}_{t|t-1} + (\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z} + (\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}_{t|t-1}), \\
&= \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t-1}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{P}^{\text{KF}}_{t|t-1}\boldsymbol{Z}' + \boldsymbol{\Sigma}_\varepsilon)^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}_{t|t-1}), \qquad \text{(A.5)}
\end{aligned}$$

where the last line, which follows from a standard matrix-inversion lemma (see equation (A.7) below), is the standard Kalman-filter level update; see e.g. Harvey (1989, p. 106) or Durbin and Koopman (2012, p. 86).

Kalman update as ESD update. We take the ESD update (2) with $\boldsymbol{H}^{\text{ex}}_t = \boldsymbol{P}^{\text{KF}}_{t|t}$ and substitute the score (A.2) to yield

$$\begin{aligned}
\boldsymbol{\theta}^{\text{ex}}_{t|t} &= \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t}\,\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}), \\
&= \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}^{\text{KF}}_{t|t}\,\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y} - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}_{t|t-1}), \\
&\overset{\text{(A.3)}}{=} \boldsymbol{\theta}_{t|t-1} + ((\boldsymbol{P}^{\text{KF}}_{t|t-1})^{-1} + \boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{Z})^{-1}\,\boldsymbol{Z}'\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}_{t|t-1}), \\
&= \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{P}_{t|t-1}\boldsymbol{Z}' + \boldsymbol{\Sigma}_\varepsilon)^{-1}(\boldsymbol{y}_t - \boldsymbol{d} - \boldsymbol{Z}\boldsymbol{\theta}_{t|t-1}), \qquad \text{(A.6)}
\end{aligned}$$

where the last line, which follows from the same matrix-inversion lemma (see equation (A.7) below), is again the standard form of Kalman's level update (see references given above).

Matrix-inversion lemmas. The above derivation relied on two matrix-inversion lemmas for positive-definite matrices $\boldsymbol{A}$, $\boldsymbol{B} \succ \mathbf{O}_k$ and an arbitrary (size-compatible) matrix $\boldsymbol{C}$:

$$(\boldsymbol{A} + \boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{C}'(\boldsymbol{B} + \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}')^{-1}. \qquad \text{(A.7)}$$

$$(\boldsymbol{A} + \boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}'(\boldsymbol{B} + \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}. \qquad \text{(A.8)}$$

While the second identity is a special case of the Woodbury matrix identity (e.g. Sherman and Morrison, 1950 and Henderson and Searle, 1981), a simple proof for the first identity is hard to find. A short proof of both identities can be found in Lange (2024b).

A.2 Lemma 1

Update stability for ISD filters

Differentiating the ISD update step (1), the Jacobian of the update with respect to the prediction is

$$\frac{\mathrm{d}\boldsymbol{\theta}^{\text{im}}_{t|t}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}} = \boldsymbol{I}_k + \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}^{\text{im}}_t\frac{\mathrm{d}\boldsymbol{\theta}^{\text{im}}_{t|t}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}}, \qquad \text{(A.9)}$$

where $\boldsymbol{\mathcal{H}}_t^{\text{im}} := \nabla^2(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}^{\text{im}})$ is the Hessian matrix evaluated at the ISD update. For clarity of exposition, from now on we omit the superscript on $\boldsymbol{\theta}_{t|t-1}$. Pre-multiply by the penalty matrix $\boldsymbol{P}$ to get

$$\boldsymbol{P}\frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{im}}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}} = \boldsymbol{P} + \boldsymbol{\mathcal{H}}_t^{\text{im}}\frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{im}}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}}, \tag{A.10}$$

and solve for the Jacobian to obtain

$$\frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{im}}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}} = (\boldsymbol{P} - \boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}\boldsymbol{P} = (\boldsymbol{I}_k - \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}. \tag{A.11}$$

Here, we note that the inverse of $\boldsymbol{P} - \boldsymbol{\mathcal{H}}_t^{\text{im}}$ exists as $\boldsymbol{P} - \boldsymbol{\mathcal{H}}_t^{\text{im}} \succ \boldsymbol{O}$ by Assumption 1, i.e. the penalty exceeds the possible non-concavity of the log density such that the regularized objective is strictly concave. The second equality follows from $(\boldsymbol{P} - \boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}\boldsymbol{P} = [\boldsymbol{P}^{-1}(\boldsymbol{P} - \boldsymbol{\mathcal{H}}_t^{\text{im}})]^{-1} = (\boldsymbol{I}_k - \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}$.

Next, we investigate

$$\begin{aligned}
\left\| \frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{im}}}{\mathrm{d}\boldsymbol{\theta}_{t|t-1}^{\text{im}\prime}} \right\|_{\boldsymbol{P}} = \left\|(\boldsymbol{I}_k - \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}\right\|_{\boldsymbol{P}} &= \left\|\boldsymbol{P}^{1/2}\,(\boldsymbol{I}_k - \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}\,\boldsymbol{P}^{-1/2}\right\| \\
&= \left\|(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2})^{-1}\right\| \\
&= \lambda_{\max}\left((\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2})^{-1}\right) \\
&= \lambda_{\min}(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2})^{-1} \\
&= \left(1 + \lambda_{\min}\left(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2}\right)\right)^{-1},
\end{aligned} \tag{A.12}$$

where the first line uses the induced (ellipsoid) matrix norm (e.g. Jungers, 2009, p. 39) $\|\boldsymbol{A}\|_{\boldsymbol{W}} = \|\boldsymbol{W}^{1/2}\boldsymbol{A}\boldsymbol{W}^{-1/2}\|$ for any symmetric positive definite matrix $\boldsymbol{W}$ and square matrix $\boldsymbol{A}$ of equal dimension, the second line uses

$$\boldsymbol{P}^{1/2}(\boldsymbol{I}_k - \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{im}})^{-1}\boldsymbol{P}^{-1/2} = \left[\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2}\right]^{-1}. \tag{A.13}$$

The third line uses that $\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2}$ is (symmetric and) positive definite, which follows from the symmetry of $\boldsymbol{P}$ along with the condition $\boldsymbol{P} - \boldsymbol{\mathcal{H}}^{\text{im}} \succ \boldsymbol{\mathrm{O}}_k$. The fourth and fifth lines use $\lambda_{\max}(\boldsymbol{A}^{-1}) = 1/\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{I}_k + \boldsymbol{A}) = 1 + \lambda_{\min}(\boldsymbol{A})$ for arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$, respectively.

Furthermore, note that $\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2} \succ \boldsymbol{\mathrm{O}}_k$ implies that $-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2} \succ -\boldsymbol{I}_k$ such that $\lambda_{\min}(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2}) > -1$. In addition, we have that the function $f(x) = (1+x)^{-1}$ is decreasing in x for $x > -1$. Together this means we need to lower bound $\lambda_{\min}(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2})$ to obtain an upper bound for $\left\| \frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{im}}}{\mathrm{d}\boldsymbol{\theta}_{t|t-1}^{\text{im}\prime}} \right\|_{\boldsymbol{P}}$. Hence, we consider

$$\lambda_{\min}(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2}) = \lambda_{\min}(-\boldsymbol{P}^{-1/2}(\boldsymbol{\mathcal{H}}_t^{\text{im}} - \alpha^-\boldsymbol{I}_k)\boldsymbol{P}^{-1/2} - \alpha^-\boldsymbol{P}^{-1})$$

$$
\begin{aligned}
&\geq \lambda_{\min}(-\boldsymbol{P}^{-1/2}(\boldsymbol{\mathcal{H}}_t^{\text{im}} - \alpha^- \boldsymbol{I}_k)\boldsymbol{P}^{-1/2}) + \lambda_{\min}(-\alpha^- \boldsymbol{P}^{-1}) \\
&= \lambda_{\min}(\boldsymbol{P}^{-1/2}(\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{im}})\boldsymbol{P}^{-1/2}) - \alpha^- \lambda_{\max}(\boldsymbol{P}^{-1}) \\
&\geq \lambda_{\min}(\boldsymbol{P}^{-1})\lambda_{\min}(\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{im}}) - \alpha^- / \lambda_{\min}(\boldsymbol{P}), \\
&= \frac{\alpha^- + \lambda_{\min}(-\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}
\end{aligned} \tag{A.14}
$$

where the second line use $\lambda_{\min}(\boldsymbol{A}+\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B})$ for symmetric matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same size. In addition, $\alpha^- := 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \geq 0$ is the concavity violation magnitude, which is 0 in case of concavity and equal to the maximum eigenvalue of $\boldsymbol{\mathcal{H}}_t^{\text{im}}$ else. Therefore, $\lambda_{\min}(\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{im}}) = \alpha^- + \lambda_{\min}(-\boldsymbol{\mathcal{H}}_t^{\text{im}}) = \alpha^- - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) = 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) = -\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})(1 - 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]) = -1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \geq 0$, such that $\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{im}} \succeq \boldsymbol{O}_k$. As a result, the fourth line above is able to use $\lambda_{\min}(\boldsymbol{ABA}) \geq \lambda_{\min}(\boldsymbol{A}^2)\lambda_{\min}(\boldsymbol{B})$ for positive semi-definite $\boldsymbol{A}, \boldsymbol{B}$ of the same size. To see this, note that $\lambda_{\min}(\boldsymbol{ABA}) = \min_{\boldsymbol{x} \neq \boldsymbol{0}_k}(\boldsymbol{x}'\boldsymbol{ABAx})/(\boldsymbol{x}'\boldsymbol{x}) \geq \min_{\boldsymbol{x} \neq \boldsymbol{0}_k} \lambda_{\min}(\boldsymbol{B})(\boldsymbol{x}'\boldsymbol{A}^2\boldsymbol{x})/(\boldsymbol{x}'\boldsymbol{x}) \geq \lambda_{\min}(\boldsymbol{B})\lambda_{\min}(\boldsymbol{A}^2)$.

Next filling in the definition of α^- and rewriting using basic properties of eigenvalues and the indicator function, from equation (A.14) we obtain

$$
\begin{aligned}
\lambda_{\min}(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2}) &\geq \frac{\alpha^- + \lambda_{\min}(-\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})} \\
&= \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})} - \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\min}(\boldsymbol{P})} \\
&= -\frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})} - \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\min}(\boldsymbol{P})},
\end{aligned} \tag{A.15}
$$

where the third line uses $1(x > 0) - 1 = -1(x \leq 0)$ for $\forall x \in \mathbb{R}$. In other words, we obtain different bounds depending on whether we are in the concave or non-concave case. Substituting equation (A.15) in our expression for the Jacobian (A.12), we obtain

$$
\begin{aligned}
\left\| \frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{im}}}{\mathrm{d}\boldsymbol{\theta}_{t|t-1}^{\text{im}\prime}} \right\|_{\boldsymbol{P}} &= \left(1 + \lambda_{\min}(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{im}}\boldsymbol{P}^{-1/2})\right)^{-1} \\
&\leq \left(1 - \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})} - \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\min}(\boldsymbol{P})}\right)^{-1} \\
&= \left(1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]\left(1 - \frac{\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})}\right) + 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\left(1 - \frac{\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\min}(\boldsymbol{P})}\right)\right)^{-1} \\
&= 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]\left(1 - \frac{\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\max}(\boldsymbol{P})}\right)^{-1} + 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\left(1 - \frac{\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}{\lambda_{\min}(\boldsymbol{P})}\right)^{-1} \\
&= 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]\left(\frac{\lambda_{\max}(\boldsymbol{P})}{\lambda_{\max}(\boldsymbol{P}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}\right) + 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]\left(\frac{\lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \max\left\{ \frac{\lambda_{\max}(\boldsymbol{P})}{\lambda_{\max}(\boldsymbol{P}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})} \,,\; \frac{\lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})} \right\} \\
&= \max\left\{ \frac{\lambda_{\max}(\boldsymbol{P})}{\lambda_{\max}(\boldsymbol{P}) + \lambda_{\min}(-\boldsymbol{\mathcal{H}}_t^{\text{im}})} \,,\; \frac{\lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P}) + \lambda_{\max}(-\boldsymbol{\mathcal{H}}_t^{\text{im}})} \right\} \\
&\leq \max\left\{ \frac{\lambda_{\max}(\boldsymbol{P})}{\lambda_{\max}(\boldsymbol{P}) + \alpha} \,,\; \frac{\lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P}) + \alpha} \right\} \\
&= 1 - \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P}) + \alpha^+} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P}) - \alpha^-},
\end{aligned} \tag{A.16}
$$

where the penultimate line uses that by Assumption 1, the negative Hessian is lower bounded $\alpha \boldsymbol{I}_k \preceq -\boldsymbol{\mathcal{H}}_t^{\text{im}}$, and last line uses that $\alpha = \alpha^+ - \alpha^-$, where $\alpha^+ := \max\{0, \alpha\}$ and $\alpha^- := \max\{0, -\alpha\}$. In sum, in the concave case ($1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) \leq 0]$), we obtain the standard contraction coefficient $\frac{\lambda_{\max}(\boldsymbol{P})}{\lambda_{\max}(\boldsymbol{P}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}$ (see also Lange et al., 2024, p.9). In the non-concave case ($1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}}) > 0]$), we obtain the maximal expansion coefficient $\frac{\lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})}$. The strict concavity of the regularized objective (Assumption 1) implies $\lambda_{\max}(\boldsymbol{P}) \geq \lambda_{\min}(\boldsymbol{P}) > \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{im}})$, such that the denominator is strictly positive in either case. This concludes the proof.

Update stability for ESD filters

Differentiating the ESD update step (2), the Jacobian of the update with respect to the prediction is

$$
\frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{ex}}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}} = \boldsymbol{I}_k + \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{ex}}, \tag{A.17}
$$

where $\boldsymbol{\mathcal{H}}_t^{\text{ex}} := \nabla^2(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}^{\text{ex}})$ is the Hessian matrix evaluated at the ESD prediction. Then

$$
\begin{aligned}
\left\| \frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{ex}}}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}} \right\|_{\boldsymbol{P}} &= \left\| \boldsymbol{I}_k + \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{ex}} \right\|_{\boldsymbol{P}} = \left\| \boldsymbol{P}^{1/2}(\boldsymbol{I}_k + \boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}_t^{\text{ex}})\boldsymbol{P}^{-1/2} \right\| \\
&= \left\| \boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2} \right\| \\
&= \max\left\{ \lambda_{\max}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}), -\lambda_{\min}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}) \right\},
\end{aligned} \tag{A.18}
$$

where the first line uses the definition of the induced matrix norm $\|\boldsymbol{A}\|_{\boldsymbol{W}} = \|\boldsymbol{W}^{1/2}\boldsymbol{A}\boldsymbol{W}^{-1/2}\|$ for any symmetric positive definite matrix $\boldsymbol{W}$ and square matrix $\boldsymbol{A}$ of equal dimension, and the second line uses $\|\boldsymbol{A}\| = \sqrt{\lambda_{\max}(\boldsymbol{A}^2)} = \max\{\lambda_{\max}(\boldsymbol{A}), -\lambda_{\min}(\boldsymbol{A})\}$ for any symmetric matrix $\boldsymbol{A}$ with real eigenvalues. Because we work exclusively with real-valued matrices, we have that symmetry is sufficient to guarantee that all eigenvalues are real.

We first consider the term associated with the maximum eigenvalue of $\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}$,

$$
\lambda_{\max}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}) = 1 + \lambda_{\max}(\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2})
$$

$$
\begin{aligned}
&= 1 + \lambda_{\max}(\boldsymbol{P}^{-1/2}(\boldsymbol{\mathcal{H}}_t^{\text{ex}} - \alpha^- \boldsymbol{I}_k)\boldsymbol{P}^{-1/2} + \alpha^- \boldsymbol{P}^{-1}) \\
&= 1 - \lambda_{\min}(\boldsymbol{P}^{-1/2}(\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{ex}})\boldsymbol{P}^{-1/2} - \alpha^- \boldsymbol{P}^{-1}) \\
&\leq 1 - \left[\lambda_{\min}(\boldsymbol{P}^{-1/2}(\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{ex}})\boldsymbol{P}^{-1/2}) + \lambda_{\min}(-\alpha^- \boldsymbol{P}^{-1})\right] \\
&\leq 1 - \lambda_{\min}(\boldsymbol{P}^{-1})\lambda_{\min}(\alpha^- \boldsymbol{I}_k - \boldsymbol{\mathcal{H}}_t^{\text{ex}}) + \alpha^- \lambda_{\max}(\boldsymbol{P}^{-1}) \\
&= 1 - \frac{\alpha^- - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\max}(\boldsymbol{P})} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})} \\
&= 1 - \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \geq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) - \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\max}(\boldsymbol{P})} + \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \geq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\min}(\boldsymbol{P})} \\
&= 1 - \frac{-1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) < 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\max}(\boldsymbol{P})} + \frac{1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \geq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\min}(\boldsymbol{P})} \\
&= 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) < 0]\left(\frac{\lambda_{\max}(\boldsymbol{P}) + \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\max}(\boldsymbol{P})}\right) + 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \geq 0]\left(\frac{\lambda_{\min}(\boldsymbol{P}) + \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\min}(\boldsymbol{P})}\right) \\
&= \max\left\{\frac{\lambda_{\max}(\boldsymbol{P}) + \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\max}(\boldsymbol{P})}, \frac{\lambda_{\min}(\boldsymbol{P}) + \lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\min}(\boldsymbol{P})}\right\} \\
&= \max\left\{\frac{\lambda_{\max}(\boldsymbol{P}) - \lambda_{\min}(-\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\max}(\boldsymbol{P})}, \frac{\lambda_{\min}(\boldsymbol{P}) - \lambda_{\min}(-\boldsymbol{\mathcal{H}}_t^{\text{ex}})}{\lambda_{\min}(\boldsymbol{P})}\right\} \\
&\leq \max\left\{\frac{\lambda_{\max}(\boldsymbol{P}) - \alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{\lambda_{\min}(\boldsymbol{P}) - \alpha}{\lambda_{\min}(\boldsymbol{P})}\right\}. \qquad \text{(A.19)}
\end{aligned}
$$

where the first line uses that $\lambda_{\max}(\boldsymbol{I}_k + \boldsymbol{A}) = 1 + \lambda_{\max}(\boldsymbol{A})$ for arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$, the second line adds and subtracts $\alpha^- \boldsymbol{P}^{-1}$, the third line that $\lambda_{\max}(\boldsymbol{A}) = -\lambda_{\min}(-\boldsymbol{A})$ for arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$, the fourth line uses $\lambda_{\min}(\boldsymbol{A}+\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B})$ for symmetric matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same size, the fifth line that $\lambda_{\min}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}) \geq \lambda_{\min}(\boldsymbol{A}^2)\lambda_{\min}(\boldsymbol{B})$ for positive semi-definite $\boldsymbol{A}, \boldsymbol{B}$ of the same size, the sixth line that $\lambda_{\min}(\boldsymbol{A}^{-1}) = 1/\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A}^{-1}) = 1/\lambda_{\min}(\boldsymbol{A})$ for positive semi-definite matrix $\boldsymbol{A}$, the seventh line that $\alpha^- := 1[\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \geq 0]\lambda_{\max}(\boldsymbol{\mathcal{H}}_t^{\text{ex}})$, the eighth line that $1(x \geq 0) - 1 = -1(x < 0)$ for $\forall x \in \mathbb{R}$. Furthermore, the ninth line uses that we may write $1 = 1(x < 0) + 1(x \geq 0) = [\lambda_{\max}(\boldsymbol{P})/\lambda_{\max}(\boldsymbol{P})]1(x < 0) + [\lambda_{\min}(\boldsymbol{P})/\lambda_{\min}(\boldsymbol{P})]1(x \geq 0)$. The penultimate line uses that $-\lambda_{\max}(\boldsymbol{A}) = \lambda_{\min}(-\boldsymbol{A})$ for arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$, and the final line uses that by Assumption 1, the negative Hessian is lower bounded $\alpha \boldsymbol{I}_k \preceq -\boldsymbol{\mathcal{H}}_t^{\text{ex}}$.

Next, we consider the term associated with the minimum eigenvalue of $\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}$,

$$
\begin{aligned}
-\lambda_{\min}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}) &= -1 - \lambda_{\min}(\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}) \\
&= -1 + \lambda_{\max}(-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_t^{\text{ex}}\boldsymbol{P}^{-1/2}) \\
&= -1 + \max_{\boldsymbol{x}\neq \boldsymbol{0}_d} \frac{\boldsymbol{x}'\boldsymbol{P}^{-1/2}(-\boldsymbol{\mathcal{H}}_t^{\text{ex}})\boldsymbol{P}^{-1/2}\boldsymbol{x}}{\|\boldsymbol{x}\|} \\
&\leq -1 + \max_{\boldsymbol{x}\neq \boldsymbol{0}_d} \lambda_{\max}(-\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \frac{\|\boldsymbol{P}^{-1/2}\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|}
\end{aligned}
$$

$$
\begin{aligned}
&\leq -1 + \beta \max_{\boldsymbol{x} \neq \mathbf{0}_d} \frac{\boldsymbol{x}' \boldsymbol{P}^{-1} \boldsymbol{x}}{\|\boldsymbol{x}\|} \\
&= -1 + \beta \lambda_{\max}(\boldsymbol{P}^{-1}) \\
&= \frac{\beta - \lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P})},
\end{aligned} \tag{A.20}
$$

where the first line uses that $\lambda_{\min}(\boldsymbol{I}_k + \boldsymbol{A}) = 1 + \lambda_{\min}(\boldsymbol{A})$ and the second line that $-\lambda_{\min}(\boldsymbol{A}) = \lambda_{\max}(-\boldsymbol{A})$ both for arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k \times k}$. The third line uses the Rayleigh quotient to express the maximum eigenvalue of a symmetric matrix, the fifth that $\lambda_{\max}(-\boldsymbol{\mathcal{H}}_t^{\text{ex}}) \leq \beta$ by Assumption 1 and that $\beta > 0$, such that it may be moved outside the max operator. The penultimate line then recognizes the remaining maximization to reflect the maximum eigenvalue of $\boldsymbol{P}^{-1}$.

Combining the eigenvalue bounds for $\boldsymbol{I}_k + \boldsymbol{P}^{-1/2} \boldsymbol{\mathcal{H}}_t^{\text{ex}} \boldsymbol{P}^{-1/2}$, (A.19)-(A.20), we obtain

$$
\begin{aligned}
\left\| \frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^{\text{ex}}}{\mathrm{d}\boldsymbol{\theta}_{t|t-1}^{\text{ex}\prime}} \right\|_{\boldsymbol{P}} &\leq \max \left\{ \max \left\{ \frac{\lambda_{\max}(\boldsymbol{P}) - \alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{\lambda_{\min}(\boldsymbol{P}) - \alpha}{\lambda_{\min}(\boldsymbol{P})} \right\}, \frac{\beta - \lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P})} \right\} \\
&= \max \left\{ \frac{\lambda_{\max}(\boldsymbol{P}) - \alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{\lambda_{\min}(\boldsymbol{P}) - \alpha}{\lambda_{\min}(\boldsymbol{P})}, \frac{\beta - \lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P})} \right\} \\
&= 1 + \max \left\{ \frac{-\alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{-\alpha}{\lambda_{\min}(\boldsymbol{P})}, \frac{\beta - 2\lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P})} \right\} \\
&= 1 - \min \left\{ \frac{\alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{\alpha}{\lambda_{\min}(\boldsymbol{P})}, \frac{2\lambda_{\min}(\boldsymbol{P}) - \beta}{\lambda_{\min}(\boldsymbol{P})} \right\} \\
&= 1 - \min \left\{ \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}, 2 - \frac{\beta}{\lambda_{\min}(\boldsymbol{P})} \right\},
\end{aligned} \tag{A.21}
$$

where the penultimate line uses that $\max\{a_i\} = -\min\{-a_i\}$, for $a_1, a_2, \ldots, a_n \in \mathbb{R}$, $n \in \mathbb{N}$, and the last line that $\alpha = \alpha^+ - \alpha^-$, where $\alpha^+ := \max\{0, \alpha\}$ and $\alpha^- := \max\{0, -\alpha\}$.

A.3 Lemma 2

Define $\gamma := \lambda_{\min}\left(\boldsymbol{P} - \boldsymbol{\Phi}' \boldsymbol{P} \boldsymbol{\Phi}\right) \in \mathbb{R}$. Since $\boldsymbol{P}$ is positive definite by assumption, we have that:

$$
\gamma = \lambda_{\min}\left(\boldsymbol{P} - \boldsymbol{\Phi}' \boldsymbol{P} \boldsymbol{\Phi}\right) \leq \lambda_{\min}(\boldsymbol{P}), \tag{A.22}
$$

Next, we note that

$$
\mathbf{0} \leq \boldsymbol{P} - \gamma \boldsymbol{I}_k - \boldsymbol{\Phi}' \boldsymbol{P} \boldsymbol{\Phi}, \tag{A.23}
$$

which we will use in the subsequent steps. Given that $\boldsymbol{P}$ is positive definite, we have

$$
\frac{1}{\lambda_{\max}(\boldsymbol{P})} \boldsymbol{P} \leq \boldsymbol{I}_k \leq \frac{1}{\lambda_{\min}(\boldsymbol{P})} \boldsymbol{P}. \tag{A.24}
$$

Consider two cases.

1. When $\gamma > 0$: Multiplying (A.24) by $-\gamma$, we obtain

$$\frac{-\gamma}{\lambda_{\max}(\boldsymbol{P})}\boldsymbol{P} \geq -\gamma \boldsymbol{I}_k \geq \frac{-\gamma}{\lambda_{\min}(\boldsymbol{P})}\boldsymbol{P}. \tag{A.25}$$

2. When $\gamma < 0$: Multiplying (A.24) by $-\gamma$, we obtain

$$\frac{-\gamma}{\lambda_{\max}(\boldsymbol{P})}\boldsymbol{P} \leq -\gamma \boldsymbol{I}_k \leq \frac{-\gamma}{\lambda_{\min}(\boldsymbol{P})}\boldsymbol{P}. \tag{A.26}$$

Combining both cases, we conclude that for all $\gamma \in \mathbb{R}$

$$-\gamma \boldsymbol{I}_k \leq -\min\left\{\frac{\gamma}{\lambda_{\min}(\boldsymbol{P})}, \frac{\gamma}{\lambda_{\max}(\boldsymbol{P})}\right\}\boldsymbol{P}, \tag{A.27}$$

which implies that

$$\boldsymbol{0} \leq \boldsymbol{P} - \gamma \boldsymbol{I}_k - \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi} \leq \left(1 - \min\left\{\frac{\gamma}{\lambda_{\min}(\boldsymbol{P})}, \frac{\gamma}{\lambda_{\max}(\boldsymbol{P})}\right\}\right)\boldsymbol{P} - \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi}. \tag{A.28}$$

This establishes the existence of a scalar $z \in \mathbb{R}$ such that $z^2\boldsymbol{P} - \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi} \geq \boldsymbol{0}$. Using the result from Jungers (2009, p. 39)

$$\|\boldsymbol{\Phi}\|_{\boldsymbol{P}} = \inf\left\{z \geq 0 : z^2\boldsymbol{P} - \boldsymbol{\Phi}'\boldsymbol{P}\boldsymbol{\Phi} \geq \boldsymbol{0}\right\}, \tag{A.29}$$

it follows that

$$\begin{aligned}\|\boldsymbol{\Phi}\|_{\boldsymbol{P}} &\leq \sqrt{1 - \min\left\{\frac{\gamma}{\lambda_{\min}(\boldsymbol{P})}, \frac{\gamma}{\lambda_{\max}(\boldsymbol{P})}\right\}} \\ &= \sqrt{1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}},\end{aligned} \tag{A.30}$$

where the second line follows from the decomposition $\gamma = \gamma^+ - \gamma^-$, with $\gamma^+ := \max\{0, \gamma\}$ and $\gamma^- := \max\{0, -\gamma\}$. Squaring both sides, which are non-negative, yields the final result

$$\|\boldsymbol{\Phi}\|_{\boldsymbol{P}}^2 \leq 1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}. \tag{A.31}$$

A.4 Theorem 1

Let $f_t^j : \boldsymbol{\Theta} \to \boldsymbol{\Theta}$ denote the update function at time t for $j \in \{\text{im}, \text{ex}\}$. For example, for the ISD update we have $\boldsymbol{\theta}_{t|t}^{\text{im}} = f_t^{\text{im}}(\boldsymbol{\theta}_{t|t-1}) = f^{\text{im}}(\boldsymbol{\theta}_{t|t-1}|\boldsymbol{y}_t, \boldsymbol{P}) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmax}}\,\{\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}) - \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\|_{\boldsymbol{P}}^2\}$. Because the proof structure is not contingent on whether the update is explicit or implicit, we suppress the filter-type superscript in the proof below and use f_t and $\boldsymbol{\theta}_{t|t}$.

Next, let $g_t^{\boldsymbol{a}} : [0,1] \to \mathbb{R}$ for some $\boldsymbol{a} \in \mathbb{R}^d$, where $g_t^{\boldsymbol{a}}(u) := \langle \boldsymbol{a}, f_t(u\boldsymbol{\theta}_{t|t-1} + (1-u)\widetilde{\boldsymbol{\theta}}_{t|t-1})\rangle$, $u \in [0,1]$ and $\boldsymbol{\theta}_{t|t-1}, \widetilde{\boldsymbol{\theta}}_{t|t-1} \in \boldsymbol{\Theta}$ are two predictions. Note that by convexity of the parameter space $\boldsymbol{\Theta}$, we have $u\boldsymbol{\theta}_{t|t-1} + (1-u)\widetilde{\boldsymbol{\theta}}_{t|t-1} \in \boldsymbol{\Theta}, \forall u \in [0,1]$, such that $g_t^{\boldsymbol{a}}$ is well defined.

For both the ISD and ESD update mappings, we have that the Jacobian of the update mapping f_t is well-defined if the Hessian of the log-likelihood exists, see again Lemma 1. In addition, if $f_t(\boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$ everywhere, then $g_t^{\boldsymbol{a}}(u)$ is differentiable in u everywhere. In sum, Assumption 1 implies $g_t^{\boldsymbol{a}}(u)$ is differentiable and hence continuous everywhere. Therefore, by the mean-value theorem, we have $\forall \boldsymbol{a} \in \mathbb{R}^d$ that $\exists\, u^\star \in [0,1]$ such that

$$\begin{aligned}
\langle \boldsymbol{a}, \boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\rangle &= \langle \boldsymbol{a}, f_t(\boldsymbol{\theta}_{t|t-1}) - f_t(\widetilde{\boldsymbol{\theta}}_{t|t-1})\rangle \\
&= g_t^{\boldsymbol{a}}(1) - g_t^{\boldsymbol{a}}(0) \\
&= \frac{\mathrm{d}g_t^{\boldsymbol{a}}(u)}{\mathrm{d}u}\Big|_{u=u^\star}(1-0) \\
&= \langle \boldsymbol{a}, \mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)(\boldsymbol{\theta}_{t|t-1} - \widetilde{\boldsymbol{\theta}}_{t|t-1})\rangle, \qquad \text{(A.32)}
\end{aligned}$$

where $\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star) := \frac{\mathrm{d}f_t}{\mathrm{d}\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t^\star}$ is the $k \times k$ Jacobian of the update function f_t evaluated at the midpoint $\boldsymbol{\theta}_t^\star := u^\star\boldsymbol{\theta}_{t|t-1} + (1-u^\star)\widetilde{\boldsymbol{\theta}}_{t|t-1}$.

Next, we use our result with $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t})/\|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}} \in \mathbb{R}^k$. This yields

$$\begin{aligned}
\|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}} &= \langle \boldsymbol{P}(\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t})/\|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}}, \boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\rangle \\
&= \langle \boldsymbol{P}(\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t})/\|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}}, \mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)(\boldsymbol{\theta}_{t|t-1} - \widetilde{\boldsymbol{\theta}}_{t|t-1})\rangle \\
&= \|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}}^{-1} \langle \boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}), \boldsymbol{P}^{1/2}\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)(\boldsymbol{\theta}_{t|t-1} - \widetilde{\boldsymbol{\theta}}_{t|t-1})\rangle \\
&\leq \|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}}^{-1}\|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}}\|\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)(\boldsymbol{\theta}_{t|t-1} - \widetilde{\boldsymbol{\theta}}_{t|t-1}\|_{\boldsymbol{P}} \\
&\leq \|\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}\|\boldsymbol{\theta}_{t|t-1} - \widetilde{\boldsymbol{\theta}}_{t|t-1}\|_{\boldsymbol{P}}, \qquad \text{(A.33)}
\end{aligned}$$

where the second line uses the mean-value theorem result above, the fourth uses the Cauchy-Schwarz inequality and $\|\boldsymbol{P}^{1/2}\boldsymbol{x}\| = \|\boldsymbol{x}\|_{\boldsymbol{P}}, \forall \boldsymbol{x} \in \mathbb{R}^d$ and the final line uses $\|\boldsymbol{A}\|_{\boldsymbol{P}} = \sup_{\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}_k\}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{P}}}{\|\boldsymbol{x}\|_{\boldsymbol{P}}}, \forall \boldsymbol{A} \in \mathbb{R}^{k \times k}$.

Substituting in the prediction step (3), we may obtain

$$\begin{aligned}
\|\boldsymbol{\theta}_{t|t} - \widetilde{\boldsymbol{\theta}}_{t|t}\|_{\boldsymbol{P}} &\leq \|\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}\|\boldsymbol{\theta}_{t|t-1} - \widetilde{\boldsymbol{\theta}}_{t|t-1}\|_{\boldsymbol{P}} \\
&\leq \|\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}\|\boldsymbol{\Phi}(\boldsymbol{\theta}_{t-1|t-1} - \widetilde{\boldsymbol{\theta}}_{t-1|t-1})\|_{\boldsymbol{P}} \\
&\leq \|\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}\|\boldsymbol{\Phi}\|_{\boldsymbol{P}}\|\boldsymbol{\theta}_{t-1|t-1} - \widetilde{\boldsymbol{\theta}}_{t-1|t-1}\|_{\boldsymbol{P}}, \qquad \text{(A.34)}
\end{aligned}$$

which indicates that the update-to-update mapping at time t is contractive in the norm $\|\cdot\|_{\boldsymbol{P}}$ if $\|\mathcal{J}_{f_t}(\boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}\|\boldsymbol{\Phi}\|_{\boldsymbol{P}} < 1$.

In Lemma 1, we derived bounds for the ISD updates (A.16) and ESD updates (A.21). Note that, with slight abuse of notation, we have $\mathcal{J}_{f_t^j}(\boldsymbol{\theta}_t^\star) := \frac{\mathrm{d}f_t^j}{\mathrm{d}\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t^\star} = \frac{\mathrm{d}\boldsymbol{\theta}_{t|t}^j}{\mathrm{d}\boldsymbol{\theta}'_{t|t-1}}\Big|_{\boldsymbol{\theta}_{t|t-1}=\boldsymbol{\theta}_t^\star}$ for $j \in \{\mathrm{im}, \mathrm{ex}\}$. Substituting in these bounds and $\|\boldsymbol{\Phi}\|_{\boldsymbol{P}} \leq \sqrt{1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}}$ from prediction-stability Lemma 2 in equation (A.34), we have that

$$\|\boldsymbol{\theta}_{t|t}^j - \widetilde{\boldsymbol{\theta}}_{t|t}^j\|_{\boldsymbol{P}} \leq \tau_j^{1/2} \|\boldsymbol{\theta}_{t-1|t-1}^j - \widetilde{\boldsymbol{\theta}}_{t-1|t-1}^j\|_{\boldsymbol{P}}, \tag{A.35}$$

$$\tau_{\mathrm{im}}^{1/2} := \sqrt{1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}} \left(1 - \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P}) + \alpha^+} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P}) - \alpha^-}\right), \tag{A.36}$$

$$\tau_{\mathrm{ex}}^{1/2} := \sqrt{1 - \frac{\gamma^+}{\lambda_{\max}(\boldsymbol{P})} + \frac{\gamma^-}{\lambda_{\min}(\boldsymbol{P})}} \left(1 - \min\left\{\frac{\alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{\alpha}{\lambda_{\min}(\boldsymbol{P})}, \frac{2\lambda_{\min}(\boldsymbol{P}) - \beta}{\lambda_{\min}(\boldsymbol{P})}\right\}\right), \tag{A.37}$$

such that the ISD and ESD update-to-update mappings at time t are contractive if $\tau_{\mathrm{im}}^{1/2} < 1$ and $\tau_{\mathrm{ex}}^{1/2} < 1$, respectively. Because $\tau_{\mathrm{im}}^{1/2}, \tau_{\mathrm{ex}}^{1/2} \geq 0$ (see again the proofs of Lemma 1 and Lemma 2), we may equivalently state these conditions as $\tau^{\mathrm{im}} < 1$ and $\tau^{\mathrm{ex}} < 1$, which are exactly the sufficient conditions that are stated in this theorem.

Since $\tau_{\mathrm{im}}^{1/2}, \tau_{\mathrm{ex}}^{1/2}, \|\cdot\| \geq 0$, squaring both sides of (A.35) and repeatidly applying it yields

$$\|\boldsymbol{\theta}_{t|t}^j - \widetilde{\boldsymbol{\theta}}_{t|t}^j\|_{\boldsymbol{P}}^2 \leq (\tau_j)^t \|\boldsymbol{\theta}_{0|0} - \widetilde{\boldsymbol{\theta}}_{0|0}\|_{\boldsymbol{P}}^2, \tag{A.38}$$

where $\boldsymbol{\theta}_{0|0}^j, \widetilde{\boldsymbol{\theta}}_{0|0}^j \in \boldsymbol{\Theta}$ are two starting points. Thus, under the (sufficient) condition $\tau^j < 1$, it follows that

$$\lim_{t\to\infty} \|\boldsymbol{\theta}_{t|t}^j - \widetilde{\boldsymbol{\theta}}_{t|t}^j\|_{\boldsymbol{P}}^2 = 0. \tag{A.39}$$

for any starting points $\boldsymbol{\theta}_{0|0}^j$ and $\widetilde{\boldsymbol{\theta}}_{0|0}^j$ and any data sequence $\{\boldsymbol{y}_t\}$, and this convergence to zero is exponentially fast. Finally, $\forall \boldsymbol{W}, \tilde{\boldsymbol{W}} \in \mathbb{R}^{k\times k}, \boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{x}\|_{\boldsymbol{W}}^2 = \boldsymbol{x}'\boldsymbol{W}\boldsymbol{x} = \boldsymbol{0}_k \Leftrightarrow \boldsymbol{u} = \boldsymbol{0}_k \Leftrightarrow \|\boldsymbol{u}\|_{\tilde{\boldsymbol{W}}}^2 = \boldsymbol{0}_k$ (norm equivalence), which implies exponential convergence of the filtered paths in the Euclidean norm $\|\cdot\|$, which concludes the proof.

A.5 Theorem 2

Here, we derive the values of a, b, c, d from Theorem 2 in terms of other quantities defined in Assumptions 1–5. Specifically, we derive a and b for the ISD and ESD updates, and c and d for the prediction steps assuming an unknown DGP and a known DGP that is a state-space model with linear state dynamics, respectively.

Preliminaries

Throughout this proof, we make use of two different expectation operators. First, we use the unconditional expectation $\mathbb{E}[\cdot]$, which acts on $\{\boldsymbol{y}_t\}$ using the true densities $\{p^0(\cdot \mid \boldsymbol{\vartheta}_t)\}$ and then on the true state path $\{\boldsymbol{\vartheta}_t\}$ using its joint density. Second, we use the conditional expectation operator that acts only on $\boldsymbol{y}_t$ given a particular state $\boldsymbol{\vartheta}_t$. That is, $\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[\cdot] := \int \cdot \, p^0(\boldsymbol{y}|\boldsymbol{\vartheta}_t) d\boldsymbol{y}$. By the tower property, it follows that $\mathbb{E}[\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[\cdot]] = \mathbb{E}[\cdot]$.

For both the ESD update and the misspecified prediction step, we make use of Young's inequality. Specifically, we use that for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^k$ and positive definite $\boldsymbol{W} \in \mathbb{R}^{k \times k}$:

$$\begin{aligned}
\|\boldsymbol{u} + \boldsymbol{v}\|_{\boldsymbol{W}}^2 &= \|\boldsymbol{W}^{1/2}\boldsymbol{u} + \boldsymbol{W}^{1/2}\boldsymbol{v}\|^2 \\
&= \|\boldsymbol{W}^{1/2}\boldsymbol{u}\|^2 + \|\boldsymbol{W}^{1/2}\boldsymbol{v}\|^2 + 2\langle \boldsymbol{W}^{1/2}\boldsymbol{u}, \boldsymbol{W}^{1/2}\boldsymbol{v}\rangle \\
&\leq \|\boldsymbol{u}\|_{\boldsymbol{W}}^2 + \|\boldsymbol{v}\|_{\boldsymbol{W}}^2 + 2|\langle \boldsymbol{W}^{1/2}\boldsymbol{u}, \boldsymbol{W}^{1/2}\boldsymbol{v}\rangle| \\
&\leq \|\boldsymbol{u}\|_{\boldsymbol{W}}^2 + \|\boldsymbol{v}\|_{\boldsymbol{W}}^2 + 2\|\boldsymbol{W}^{1/2}\boldsymbol{u}\|\|\boldsymbol{W}^{1/2}\boldsymbol{v}\| \\
&\leq \|\boldsymbol{u}\|_{\boldsymbol{W}}^2 + \|\boldsymbol{v}\|_{\boldsymbol{W}}^2 + \epsilon^2\|\boldsymbol{W}^{1/2}\boldsymbol{u}\|^2 + (1/\epsilon^2)\|\boldsymbol{W}^{1/2}\boldsymbol{v}\|^2 \\
&= (1+\epsilon^2)\|\boldsymbol{u}\|_{\boldsymbol{W}}^2 + (1+1/\epsilon^2)\|\boldsymbol{v}\|_{\boldsymbol{W}}^2,
\end{aligned} \tag{A.40}$$

for any $\epsilon > 0$, where in the fourth line we used the Cauchy-Schwartz inequality and in the fifth line Young's inequality for products.

For the misspecified prediction step, we also use the Cauchy-Schwartz inequality. For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^k$, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{u} + \boldsymbol{v}\|^2 &= \mathbb{E}[\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 + 2\langle \boldsymbol{u}, \boldsymbol{v}\rangle] \\
&\leq \mathbb{E}[\|\boldsymbol{u}\|^2] + \mathbb{E}[\|\boldsymbol{v}\|^2] + 2\mathbb{E}[|\langle \boldsymbol{u}, \boldsymbol{v}\rangle|] \\
&\leq \mathbb{E}[\|\boldsymbol{u}\|^2] + \mathbb{E}[\|\boldsymbol{v}\|^2] + 2\mathbb{E}[\|\boldsymbol{u}\|\|\boldsymbol{v}\|] \\
&= \mathbb{E}[\|\boldsymbol{u}\|^2] + \mathbb{E}[\|\boldsymbol{v}\|^2] + 2\mathbb{E}[\|\boldsymbol{u}\|\|\|\boldsymbol{u}\|\|] \\
&\leq \mathbb{E}[\|\boldsymbol{u}\|^2] + \mathbb{E}[\|\boldsymbol{v}\|^2] + 2\sqrt{\mathbb{E}[\|\boldsymbol{u}\|^2\|]\mathbb{E}[\|\boldsymbol{u}\|^2\|]} \\
&= \left(\sqrt{\mathbb{E}[\|\boldsymbol{u}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{v}\|^2]}\right)^2,
\end{aligned} \tag{A.41}$$

where the third line uses the Cauchy-Schwarz identity and the penultimate line uses the Cauchy-Schwarz inequality for random variables: $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$ for scalar-valued random variables X, Y.

MSE contraction for ISD update step

The first-order condition of the ISD update, for static penalty matrix $\boldsymbol{P}$ reads:

$$\boldsymbol{\theta}_{t|t} = \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}_t^{-1}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}). \tag{A.42}$$

We move $\boldsymbol{P}^{-1}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t})$ to the left-hand side, pre-multiply both sides by the symmetric square root of the penalty matrix, denoted $\boldsymbol{P}^{\frac{1}{2}}$, and subtract $\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{\theta}_t^\star - \boldsymbol{P}^{-\frac{1}{2}}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)$ from both sides

$$\boldsymbol{P}^{\frac{1}{2}}(\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}_t^\star)-\boldsymbol{P}^{-\frac{1}{2}}(\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t})-\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)) = \boldsymbol{P}^{\frac{1}{2}}(\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}_t^\star)+\boldsymbol{P}^{-\frac{1}{2}}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star). \quad \text{(A.43)}$$

Using Rieman integrability of the Hessian of the log-likelihood function (Assumption 1), we may write

$$\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star) = \boldsymbol{\mathcal{H}}_{t|t}^\star(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star), \quad \text{(A.44)}$$

where $\boldsymbol{\mathcal{H}}_{t|t}^\star := \int_0^1 \frac{\partial^2\ell(\boldsymbol{y}_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=u\,\boldsymbol{\theta}_{t|t}+(1-u)\,\boldsymbol{\theta}_t^\star} \mathrm{d}u$ is the average Hessian between $\boldsymbol{\theta}_{t|t}$ and $\boldsymbol{\theta}_t^\star$. Substituting this result into Equation (A.43) produces

$$(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})\boldsymbol{P}^{\frac{1}{2}}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star) = \boldsymbol{P}^{\frac{1}{2}}(\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star) + \boldsymbol{P}^{-\frac{1}{2}}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star), \quad \text{(A.45)}$$

where by Assumption 2, $\boldsymbol{P} \succ \boldsymbol{\mathcal{H}}_{t|t}^\star \Rightarrow \boldsymbol{I}_k \succ \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2} \Rightarrow \boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2} \succ \mathbf{O}_k$, such that taking the inner product on both sides gives

$$\|\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}_t^\star)\|^2_{(\boldsymbol{I}_k-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2} = \|\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}}+\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|^2_{\boldsymbol{P}^{-1}}+2\langle\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}_t^\star, \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\rangle. \quad \text{(A.46)}$$

Next, we take the unconditional expectation on both sides and use that $\mathbb{E}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|^2_{\boldsymbol{P}^{-1}}] \leq \lambda_{\max}(\boldsymbol{P}^{-1})\mathbb{E}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\|^2_{\boldsymbol{P}^{-1}}] \leq \sigma^2/\lambda_{\min}(\boldsymbol{P})$ by Assumption 4(ii) and that $\mathbb{E}[\langle\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\rangle] = \mathbb{E}[\mathbb{E}_{\boldsymbol{y}_t}[\langle\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)\rangle]] = \mathbb{E}[\langle\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star, \mathbb{E}_{\boldsymbol{y}_t}[\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)]\rangle] = 0$ by the tower property and the fact that $\mathbb{E}_{\boldsymbol{y}_t}[\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t^\star)] = 0$ because the pseudo-truth uniquely minimizes the expected log likelihood (see Definition 4). This produces

$$\mathbb{E}[\|\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star)\|^2_{(\boldsymbol{I}_k-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2}] \leq \mathbb{E}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}}] + \sigma^2/\lambda_{\min}(\boldsymbol{P}). \quad \text{(A.47)}$$

For the left-hand side, we may use the following lower bound

$$\lambda_{\min}((\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2)\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}} \leq \|\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star)\|^2_{(\boldsymbol{I}_k-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2}, \quad \text{(A.48)}$$

as $\boldsymbol{I}_k-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2} \succ \mathbf{O}_k$ is positive definite, such that $\lambda_{\min}((\boldsymbol{I}_k-\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2) = \lambda_{\min}(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2 > 0$. Using the lower bound in (A.47) and dividing both sides by $\lambda_{\min}(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^2$ gives

$$\mathbb{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}}] \leq \lambda_{\min}(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^{-2}\left(\mathbb{E}[\|\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}_t^\star\|^2_{\boldsymbol{P}}] + \sigma^2/\lambda_{\min}(\boldsymbol{P})\right). \quad \text{(A.49)}$$

Using the same methodology as in the proof of Lemma 1, equation (A.16), we may obtain:

$$\lambda_{\min}(\boldsymbol{I}_k - \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}_{t|t}^\star\boldsymbol{P}^{-1/2})^{-1} \leq 1 - \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P}) + \alpha^+} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P}) - \alpha^-}. \quad \text{(A.50)}$$

As both sides are nonnegative, we may square both sides. Using the definition of the $\boldsymbol{P}$-weighted MSE produces the final result:

$$\mathrm{MSE}^{\boldsymbol{P}}_{t|t} \leq a\mathrm{MSE}^{\boldsymbol{P}}_{t|t-1} + b, \tag{A.51}$$

where $a = \left(1 - \frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})+\alpha^+} + \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})-\alpha^-}\right)^2$ and $b = a\sigma^2/\lambda_{\min}(\boldsymbol{P})$, which is the first line of the table in Theorem 2.

MSE contraction for ESD update step

The ESD update reads:

$$\boldsymbol{\theta}_{t|t} = \boldsymbol{\theta}_{t|t-1} + \boldsymbol{P}^{-1}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}), \tag{A.52}$$

where pre-multiplying both sides with $\boldsymbol{P}^{1/2}$ and subtracting $\boldsymbol{P}^{1/2}\boldsymbol{\theta}^\star_t$ on both sides yields

$$\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}^\star_t) = \boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t) + \boldsymbol{P}^{-1/2}\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}). \tag{A.53}$$

Taking the inner product and afterwards the expectation produces

$$\mathbb{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{P}}] = \mathbb{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{P}}] + 2\mathbb{E}[\langle\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\rangle] + \mathbb{E}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|^2_{\boldsymbol{P}^{-1}}]. \tag{A.54}$$

For the second term on the right-hand side of (A.54), we may write

$$\begin{aligned}
2\mathbb{E}[\langle\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\rangle] &= 2\mathbb{E}[\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[\langle\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\rangle]] \\
&= 2\mathbb{E}[\langle\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)], \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\rangle] \\
&= 2\mathbb{E}[\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[\langle\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\rangle]] \\
&= 2\mathbb{E}[\langle\boldsymbol{\mathcal{H}}^\star_{t|t-1}(\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t), \boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t\rangle] \\
&= \mathbb{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}^\star_t\|^2_{2\boldsymbol{\mathcal{H}}^\star_{t|t-1}}],
\end{aligned} \tag{A.55}$$

where the first and fourth line use the tower property, the second that $\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t] = 0$ by definition of the pseudo-truth $\boldsymbol{\theta}^\star_t$ and the fourth that

$$\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t) = \boldsymbol{\mathcal{H}}^\star_{t|t-1}(\boldsymbol{\theta}_{t|t-1} - \boldsymbol{\theta}^\star_t), \tag{A.56}$$

where $\boldsymbol{\mathcal{H}}^\star_{t|t-1} := \int_0^1 \frac{\partial^2\ell(\boldsymbol{y}_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=u\,\boldsymbol{\theta}_{t|t-1}+(1-u)\,\boldsymbol{\theta}^\star_t} \mathrm{d}u$ the average Hessian between $\boldsymbol{\theta}_{t|t-1}$ and $\boldsymbol{\theta}^\star_t$.

For the final term on the right-hand side of (A.54), we have that

$$\begin{aligned}
\mathbb{E}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1})\|^2_{\boldsymbol{P}^{-1}}] &= \mathbb{E}\left[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t) + \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1}}\right] \\
&\leq \mathbb{E}\left[(1+\chi^2)\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1}} + (1+1/\chi^2)\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1}}\right]
\end{aligned}$$

$$= (1+\chi^2)\mathbb{E}\left[\|\boldsymbol{\mathcal{H}}^\star_{t|t-1}(\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1}}\right] + (1+1/\chi^2)\mathbb{E}\left[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1}}\right]$$
$$\leq (1+\chi^2)\mathbb{E}\left[\|\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}^\star_{t|t-1}}\right] + (1+1/\chi^2)\sigma^2/\lambda_{\min}(\boldsymbol{P}), \tag{A.57}$$

where the second line uses Young's inequality as specified in (A.40) with $\epsilon = \chi$, $\boldsymbol{u} = \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}_{t|t-1}) - \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)$, $\boldsymbol{v} = \nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)$ and $\boldsymbol{W} = \boldsymbol{P}^{-1}$, the third uses the definition of $\boldsymbol{\mathcal{H}}^\star_{t|t-1}$ and the fourth that $\mathbb{E}[\|\nabla\ell(\boldsymbol{y}_t \mid \boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1}}] \leq \sigma^2/\lambda_{\min}(\boldsymbol{P})$ by Assumption 4(ii).

Together, this yields

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{P}}] \\
&\leq \mathbb{E}[\|\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{P}}] + \mathbb{E}[\|\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}^\star_t\|^2_{2\boldsymbol{\mathcal{H}}^\star_{t|t-1}}] + (1+\chi^2)\mathbb{E}\left[\|\boldsymbol{\theta}_{t|t-1}-\boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}^\star_{t|t-1}}\right] \\
&\qquad + \frac{(1+1/\chi^2)\sigma^2}{\lambda_{\min}(\boldsymbol{P})} \\
&= \mathbb{E}[\|\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}^\star_t\|^2_{\boldsymbol{P}+2\boldsymbol{\mathcal{H}}^\star_{t|t-1}+(1+\chi^2)\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}^\star_{t|t-1}}] + \frac{(1+1/\chi^2)\sigma^2}{\lambda_{\min}(\boldsymbol{P})} \\
&= \mathbb{E}[\|\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}^\star_t)\|^2_{(\boldsymbol{I}_k+\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2})^2} + \frac{(1+1/\chi^2)\sigma^2}{\lambda_{\min}(\boldsymbol{P})} \\
&+ \chi^2\mathbb{E}[\|\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t|t}-\boldsymbol{\theta}^\star_t)\|^2_{\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2}}.
\end{aligned} \tag{A.58}$$

Using the eigenvalue bounds for $\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^{\text{ex}}_t\boldsymbol{P}^{-1/2}$ in equation (A.21) from Lemma 1, we may square both sides as they are both positive semi-definite, to obtain

$$\lambda_{\max}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2})^2 \leq \left(1 - \min\left\{\frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}, 2 - \frac{\beta}{\lambda_{\min}(\boldsymbol{P})}\right\}\right)^2. \tag{A.59}$$

Similarly, we have that

$$\begin{aligned}
&\lambda_{\max}(\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2}) = \lambda_{\max}((\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2})^2) \\
&\quad = \max\{\lambda_{\max}(\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2}), -\lambda_{\min}(\boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2})\}^2 \\
&\quad = \max\{\lambda_{\max}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2}) - 1, 1 - \lambda_{\min}(\boldsymbol{I}_k + \boldsymbol{P}^{-1/2}\boldsymbol{\mathcal{H}}^\star_{t|t-1}\boldsymbol{P}^{-1/2})\}^2 \\
&\quad \leq \max\left\{\max\left\{\frac{\lambda_{\max}(\boldsymbol{P})-\alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{\lambda_{\min}(\boldsymbol{P})-\alpha}{\lambda_{\min}(\boldsymbol{P})}\right\} - 1, 1 + \frac{\beta - \lambda_{\min}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P})}\right\}^2 \\
&\quad \leq \max\left\{\frac{-\alpha}{\lambda_{\max}(\boldsymbol{P})}, \frac{-\alpha}{\lambda_{\min}(\boldsymbol{P})}, \frac{\beta}{\lambda_{\min}(\boldsymbol{P})}\right\}^2 \\
&\quad \leq \max\left\{\frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}, \frac{\beta}{\lambda_{\min}(\boldsymbol{P})}\right\}^2 = \frac{L^2}{\lambda_{\min}(\boldsymbol{P})^2}.
\end{aligned} \tag{A.60}$$

using $\beta > 0$, $\alpha^-/\lambda_{\min}(\boldsymbol{P}) \geq \alpha^-/\lambda_{\max}(\boldsymbol{P})$, and $L := \max\{\alpha^-, \beta\}$. Combining these results, we obtain the final result:

$$\text{MSE}^{\boldsymbol{P}}_{t|t} \leq a\text{MSE}^{\boldsymbol{P}}_{t|t-1} + b, \tag{A.61}$$

where $a = \left(1 - \min\left\{\frac{\alpha^+}{\lambda_{\max}(\boldsymbol{P})} - \frac{\alpha^-}{\lambda_{\min}(\boldsymbol{P})}, 2 - \frac{\beta}{\lambda_{\min}(\boldsymbol{P})}\right\}\right)^2 + \frac{\chi^2 L^2}{\lambda_{\min}(\boldsymbol{P})^2}$ and $b = \frac{(1+1/\chi^2)\sigma^2}{\lambda_{\min}(\boldsymbol{P})}$, which is the second line of the table in Theorem 2.

Prediction step with unknown DGP (Assumption 4)

We start by subtracting the pseudo-true state $\boldsymbol{\theta}_t^\star$ from the prediction step in equation (3), multiplying by $\boldsymbol{P}^{1/2}$ and taking the squared norm:

$$\begin{aligned}
\|\boldsymbol{\theta}_{t+1|t} - \boldsymbol{\theta}_{t+1}^\star\|_{\boldsymbol{P}}^2 &= \left\|(\boldsymbol{I}_k - \boldsymbol{\Phi})\boldsymbol{\omega} + \boldsymbol{\Phi}\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_{t+1}^\star\right\|_{\boldsymbol{P}}^2 \\
&= \left\|\boldsymbol{\Phi}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star) + (\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}) + (\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star)\right\|_{\boldsymbol{P}}^2 \\
&\leq (1+\epsilon^2)\|\boldsymbol{\Phi}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}^2 \\
&\qquad + \left(1 + \frac{1}{\epsilon^2}\right)\|(\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}) + (\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star)\|_{\boldsymbol{P}}^2, \qquad \text{(A.62)}
\end{aligned}$$

where in the second line we added and subtracted $(\boldsymbol{\Phi} - \boldsymbol{I}_k)\boldsymbol{\theta}_t^\star$ and the third uses Young's inequality as specified in (A.40) with $\boldsymbol{u} = \boldsymbol{\Phi}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star)$, $\boldsymbol{v} = (\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}) + (\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star)$ and $\boldsymbol{W} = \boldsymbol{P}$. Furthermore, we note that $\|\boldsymbol{\Phi}(\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star)\|_{\boldsymbol{P}}^2 = \|\boldsymbol{\theta}_{t|t} - \boldsymbol{\theta}_t^\star\|_{\boldsymbol{\Phi}\boldsymbol{P}\boldsymbol{\Phi}}^2$. Using submultiplicativity of the $\boldsymbol{P}$-weighted matrix norm, which follows from its definition

$$\|\boldsymbol{\Phi}\|_{\boldsymbol{P}}^2 := \sup_{\boldsymbol{y}\in\mathbb{R}^k\setminus\{\boldsymbol{0}_k\}} \frac{\|\boldsymbol{\Phi}\boldsymbol{y}\|_{\boldsymbol{P}}^2}{\|\boldsymbol{y}\|_{\boldsymbol{P}}^2} \geq \frac{\|\boldsymbol{\Phi}\boldsymbol{x}\|_{\boldsymbol{P}}^2}{\|\boldsymbol{x}\|_{\boldsymbol{P}}^2}, \quad \forall \boldsymbol{x} \neq \boldsymbol{0}_k \Rightarrow \|\boldsymbol{\Phi}\boldsymbol{x}\|_{\boldsymbol{P}}^2 \leq \|\boldsymbol{\Phi}\|_{\boldsymbol{P}}^2\|\boldsymbol{x}\|_{\boldsymbol{P}}^2, \forall \boldsymbol{x} \neq \boldsymbol{0}_k, \tag{A.63}$$

and taking the unconditional expectation of (A.62) yields

$$\text{MSE}_{t+1|t}^{\boldsymbol{P}} \leq (1+\epsilon^2)\|\boldsymbol{\Phi}\|_{\boldsymbol{P}}^2\text{MSE}_{t|t}^{\boldsymbol{P}} + \left(1 + \frac{1}{\epsilon^2}\right)\mathbb{E}[\|(\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}) + (\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star)\|_{\boldsymbol{P}}^2], \tag{A.64}$$

where $\|\boldsymbol{\Phi}\|_{\boldsymbol{P}}^2$ is the squared matrix norm of $\boldsymbol{\Phi}$ induced by the vector norm $\|\tilde{\boldsymbol{u}}\|_{\boldsymbol{P}}^2$ for $\tilde{\boldsymbol{u}} \in \mathbb{R}^k$.

Using preliminary result (A.41) with $\boldsymbol{u} = (\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega})$ and $\boldsymbol{v} = \boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|(\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}) + (\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star)\|_{\boldsymbol{P}}^2] \\
&\quad\leq \lambda_{\max}(\boldsymbol{P})\left(\sqrt{\mathbb{E}[\|(\boldsymbol{\Phi} - \boldsymbol{I}_k)(\boldsymbol{\theta}_t^\star - \boldsymbol{\omega})\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star\|^2]}\right)^2 \\
&\quad\leq \lambda_{\max}(\boldsymbol{P})\left(\sqrt{\mathbb{E}[\|\boldsymbol{\Phi} - \boldsymbol{I}_k\|^2\|\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star\|^2]}\right)^2 \\
&\quad\leq \lambda_{\max}(\boldsymbol{P})\left(\|\boldsymbol{I}_k - \boldsymbol{\Phi}\|\sqrt{\mathbb{E}[\|\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star\|^2]}\right)^2 \\
&\quad\leq \lambda_{\max}(\boldsymbol{P})\left(\|\boldsymbol{I}_k - \boldsymbol{\Phi}\|s_\omega + q\right)^2, \qquad \text{(A.65)}
\end{aligned}$$

where the third lines uses that the definition of induced matrix norm and the final line that $s_\omega^2 := \mathbb{E}[\|\boldsymbol{\theta}_t^\star - \boldsymbol{\omega}\|^2]$ and $\mathbb{E}[\|\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star\|^2] = \mathbb{E}[\text{tr}((\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star)(\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star)')] = \text{tr}(\mathbb{E}[(\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star)(\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star)']) \leq \text{tr}(\boldsymbol{Q}) =: q^2 < \infty$.

Combining this result with (A.64) yields the final result

$$\text{MSE}_{t+1|t}^{\boldsymbol{P}} \leq c\text{MSE}_{t|t}^{\boldsymbol{P}} + d, \tag{A.66}$$

where $c = (1+\epsilon^2)\|\boldsymbol{\Phi}\|_{\boldsymbol{P}}^2$ and $d = \left(1+\frac{1}{\epsilon^2}\right)\lambda_{\max}(\boldsymbol{P})\left(\|\boldsymbol{I}_k - \boldsymbol{\Phi}\|s_\omega + q\right)^2$, which is the third line of the table in Theorem 2.

Prediction step with known DGP (Assumption 5)

When Assumption 5) holds, we have

$$\begin{aligned}
\text{MSE}_{t+1|t}^{\boldsymbol{P}} &= \mathbb{E}[\|\boldsymbol{\theta}_{t+1|t} - \boldsymbol{\vartheta}_{t+1}\|_{\boldsymbol{P}}^2] \\
&= \mathbb{E}[\|(\boldsymbol{I}_k - \boldsymbol{\Phi})\boldsymbol{\omega} + \boldsymbol{\Phi}\boldsymbol{\theta}_{t|t} - (\boldsymbol{I}_k - \boldsymbol{\Phi}_0)\boldsymbol{\omega}_0 - \boldsymbol{\Phi}_0\boldsymbol{\vartheta}_t - \boldsymbol{\xi}_{t+1}\|_{\boldsymbol{P}}^2] \\
&= \mathbb{E}[\|\boldsymbol{\Phi}_0(\boldsymbol{\theta}_{t|t} - \boldsymbol{\vartheta}_t) - \boldsymbol{\xi}_{t+1}\|_{\boldsymbol{P}}^2] \\
&= \mathbb{E}[\|\boldsymbol{\Phi}_0(\boldsymbol{\theta}_{t|t} - \boldsymbol{\vartheta}_t)\|_{\boldsymbol{P}}^2] + \mathbb{E}[\|\boldsymbol{\xi}_{t+1}\|_{\boldsymbol{P}}^2] - 2\mathbb{E}[\langle \boldsymbol{P}\boldsymbol{\Phi}_0(\boldsymbol{\theta}_{t|t} - \boldsymbol{\vartheta}_t), \boldsymbol{\xi}_{t+1}\rangle] \\
&= \mathbb{E}[\|\boldsymbol{\Phi}_0(\boldsymbol{\theta}_{t|t} - \boldsymbol{\vartheta}_t)\|_{\boldsymbol{P}}^2] + \mathbb{E}[\|\boldsymbol{\xi}_{t+1}\|_{\boldsymbol{P}}^2] \\
&\leq \|\boldsymbol{\Phi}_0\|_{\boldsymbol{P}}^2\mathbb{E}[\|\boldsymbol{\theta}_{t|t} - \boldsymbol{\vartheta}_t\|_{\boldsymbol{P}}^2] + \lambda_{\max}(\boldsymbol{P})\mathbb{E}[\|\boldsymbol{\xi}_{t+1}\|^2 \\
&\leq \|\boldsymbol{\Phi}_0\|_{\boldsymbol{P}}^2\text{MSE}_{t|t}^{\boldsymbol{P}} + \lambda_{\max}(\boldsymbol{P})\sigma_\xi^2,
\end{aligned} \tag{A.67}$$

where the third line uses that $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}_0$, the fifth that $\boldsymbol{\xi}_{t+1}$ is independent of $\boldsymbol{\Phi}_0(\boldsymbol{\theta}_{t|t} - \boldsymbol{\vartheta}_t)$ and with $\mathbb{E}[\boldsymbol{\xi}_{t+1}] = \boldsymbol{0}_d$, the sixth the definition of the induced matrix norm and the final line that $\mathbb{E}[\|\boldsymbol{\xi}_{t+1}\|^2] = \mathbb{E}[\text{tr}(\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}_{t+1}')] = \text{tr}(\mathbb{E}[\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}_{t+1}']) = \text{tr}(\boldsymbol{\Sigma}_\xi) = \sigma_\xi^2 \leq \infty$. We conclude that $c = \|\boldsymbol{\Phi}_0\|_{\boldsymbol{P}}^2$ and $d = \lambda_{\max}(\boldsymbol{P})\sigma_\xi^2$.

A.6 Corollary 1

Computing the bound

Consider the ISD filter in case Assumption 5 does not hold. In addition, let $\tau_{\text{im}} = ac/(1+\epsilon^2) < 1$ (see Equation (11)) and define $\tilde{d} := d/(1+\frac{1}{\epsilon^2})$. Because we must ensure $a\,c < 1$ to guarantee finite asymptotic MSE bounds, it is convenient to take $\epsilon^2 := (\kappa(1-\tau_{\text{im}}))/\tau_{\text{im}}$, such that $ac = \tau_{\text{im}}(1+\epsilon^2) = \tau_{\text{im}} + \kappa(1-\tau_{\text{im}}) < 1$, $\forall \kappa \in (0,1)$. Intuitively, κ controls the extent to which we close the "gap" between τ_{im} and unity; i.e., the denominator in the asymptotic MSE bound reads $1-ac = 1-\tau_{\text{im}}-\kappa(1-\tau_{\text{im}}) = 1-\tau_{\text{im}}-\kappa+\kappa\tau_{\text{im}} = (1-\tau_{\text{im}})(1-\kappa) \in (0,1)$.

The asymptotic MSE bound $\forall \kappa \in (0,1)$ is given by

$$\begin{aligned}
\limsup_{t\to\infty} \mathrm{MSE}^{\boldsymbol{P}}_{t|t} &\leq \frac{b+ad}{1-ac} \\
&= \frac{b+a\tilde{d}(1+\frac{1}{\epsilon^2})}{1-\tau_{\mathrm{im}}(1+\epsilon^2)} \\
&= \frac{b+a\tilde{d}(1+\frac{\tau_{\mathrm{im}}}{\kappa(1-\tau_{\mathrm{im}})})}{(1-\tau_{\mathrm{im}})(1-\kappa)} \\
&= \frac{b+a\tilde{d}}{1-\tau_{\mathrm{im}}}\frac{1}{1-\kappa} + \frac{\tau_{\mathrm{im}}a\tilde{d}}{(1-\tau_{\mathrm{im}})^2}\frac{1}{\kappa(1-\kappa)} \\
&= A\frac{1}{1-\kappa} + B\frac{1}{\kappa(1-\kappa)},
\end{aligned} \tag{A.68}$$

where the second line uses the definition of $\tilde{d}$ and τ_{im} and the third line uses our specification of ϵ^2 in terms of κ. From the final line, we that $A, B \geq 0$ are given as

$$A := \frac{b+a\tilde{d}}{1-\tau_{\mathrm{im}}}, \quad B := \frac{\tau_{\mathrm{im}}a\tilde{d}}{(1-\tau_{\mathrm{im}})^2}. \tag{A.69}$$

For future reference, we note that

$$\frac{A}{B} = \frac{\frac{b+a\tilde{d}}{1-\tau_{\mathrm{im}}}}{\frac{\tau_{\mathrm{im}}a\tilde{d}}{(1-\tau_{\mathrm{im}})^2}} = \frac{1-\tau_{\mathrm{im}}}{\tau_{\mathrm{im}}}\frac{b+a\tilde{d}}{a\tilde{d}}. \tag{A.70}$$

Minimizing the bound

In case $\tau_{\mathrm{im}} = 0$, we have that $\epsilon = \infty$ minimizes the MSE bound, as can be seen in the second line of (A.68). We therefore proceed with the case $\tau_{\mathrm{im}} > 0$, which implies $B > 0$ as $\tau_{\mathrm{im}} = ac/(1+\epsilon^2) \neq 0 \Rightarrow a \neq 0$ and $q > 0 \Rightarrow \tilde{d} > 0$; see (A.69). The final line of (A.68) illustrates that in this case $\kappa \in (0,1)$ should not approach either boundary too closely.

Minimizing the bound (A.68) with respect to κ yields the following first-order condition:

$$0 = \frac{A}{(1-\kappa)^2} + \frac{B(2\kappa-1)}{\kappa^2(1-\kappa)^2}, \tag{A.71}$$

where we may multiply by $\kappa^2(1-\kappa)^2 \in (0,1)$ to obtain

$$0 = A\,\kappa^2 + 2B\,\kappa - B, \tag{A.72}$$

which is a quadratic equation in κ, the solution of which reads

$$\kappa_{\pm} = \frac{-2B \pm \sqrt{4B^2+4AB}}{2A}. \tag{A.73}$$

Because A and B are positive, we have $\kappa_+ > 0$ and $\kappa_- < 0$. Only the positive solution is of interest. Specifically, we have

$$\begin{aligned}
\kappa_\star = \kappa_+ &= \frac{-2B + \sqrt{4B^2 + 4AB}}{2A} \\
&= \frac{-2B + \sqrt{4B^2 + 4AB}}{2A} \times \underbrace{\frac{2B + \sqrt{4B^2 + 4AB}}{2B + \sqrt{4B^2 + 4AB}}}_{=1} \\
&= \frac{4AB}{4AB + 2A\sqrt{4B^2 + 4AB}} \\
&= \frac{4AB}{4AB + \sqrt{(4AB)^2 + 16A^3B}} \\
&= \frac{1}{1 + \sqrt{1 + A/B}},
\end{aligned} \tag{A.74}$$

which shows that $\kappa_\star \in (0, 1/2)$.

Using the expression for A/B in (A.70) and that $\epsilon = \kappa \frac{1-\tau_{\text{im}}}{\tau_{\text{im}}}$, we obtain

$$\begin{aligned}
\epsilon^{\star 2} &= \frac{1-\tau_{\text{im}}}{\tau_{\text{im}}} \kappa_\star = \frac{1-\tau_{\text{im}}}{\tau_{\text{im}}} \frac{1}{1 + \sqrt{1 + A/B}} \\
&= \frac{1-\tau_{\text{im}}}{\tau_{\text{im}} + \sqrt{\tau_{\text{im}}^2 + \tau_{\text{im}}(1-\tau_{\text{im}}) \frac{b + a\tilde{d}}{a\tilde{d}}}} \\
&= \frac{1-\tau_{\text{im}}}{\tau_{\text{im}} + \sqrt{\tau_{\text{im}}^2 + \tau_{\text{im}}(1-\tau_{\text{im}})(\frac{b}{a\tilde{d}} + 1)}} \\
&= \frac{1-\tau_{\text{im}}}{\tau_{\text{im}} + \sqrt{\tau_{\text{im}}^2 + \tau_{\text{im}}(1-\tau_{\text{im}}) \frac{b}{a\tilde{d}} + \tau_{\text{im}}(1-\tau_{\text{im}})}} \\
&= \frac{1-\tau_{\text{im}}}{\tau_{\text{im}} + \sqrt{\tau_{\text{im}} + \tau_{\text{im}}(1-\tau_{\text{im}}) \frac{b}{a\tilde{d}}}} \\
&= \frac{1-\tau_{\text{im}}}{\tau_{\text{im}} + \sqrt{\tau_{\text{im}} + \tau_{\text{im}}(1-\tau_{\text{im}}) \frac{\sigma^2}{\lambda_{\max}(\boldsymbol{P})\lambda_{\min}(\boldsymbol{P})(\|\boldsymbol{I}_k - \boldsymbol{\Phi}\|_2 \, s_\omega + q)^2}}},
\end{aligned} \tag{A.75}$$

where the final line uses the values of b and d for the ISD filter from Theorem 2. Note that coerciviness (the bounds tend to infinity as κ approaches his bounds) and continuity imply that this unique stationary point is, in fact, a global minimum.

A.7 Corollary 2

Using the result of Theorem 2 and that $\boldsymbol{P} = \rho \boldsymbol{I}_k$, we have that $a = \left(\frac{\rho}{\rho + \alpha}\right)^2 \in (0, 1)$ and $b = a\rho^{-1}\sigma^2$ for the ISD filter with strong concavity ($\alpha > 0$). In addition, when the DGP is a known state-space model (Assumption 5 holds) and again using $\boldsymbol{P} = \rho \boldsymbol{I}_k$, we have that

$c = \|\boldsymbol{\Phi}_0\|_{\boldsymbol{P}}^2 = \|\boldsymbol{P}^{1/2}\boldsymbol{\Phi}_0\boldsymbol{P}^{-1/2}\|^2 = \|\rho^{1/2}\boldsymbol{I}_k\boldsymbol{\Phi}_0\rho^{-1/2}\boldsymbol{I}_k\|^2 = \|\boldsymbol{\Phi}_0\|^2$ and $d = \rho\sigma_\xi^2$. Filling these into the asymptotic filter $\boldsymbol{P}$-weighted MSE bound in (19) and converting to the unweighted MSE bound, we obtain

$$\begin{aligned}\limsup_{t\to\infty}\ \mathrm{MSE}_{t|t} &= \frac{1}{\rho}\limsup_{t\to\infty}\ \mathrm{MSE}_{t|t}^{\boldsymbol{P}} \leq \frac{1}{\rho}\frac{b+ad}{1-ac} = \frac{a\sigma^2/\rho^2 + a\sigma_\xi^2}{1-a\|\boldsymbol{\Phi}_0\|^2} \\ &= \frac{\sigma^2/\rho^2+\sigma_\xi^2}{a^{-1}-\|\boldsymbol{\Phi}_0\|^2} = \frac{\sigma^2/\rho^2+\sigma_\xi^2}{\left(\frac{\rho+\alpha}{\rho}\right)^2-\|\boldsymbol{\Phi}_0\|^2} = \frac{\sigma^2+\rho^2\sigma_\xi^2}{(\rho+\alpha)^2-\rho^2\|\boldsymbol{\Phi}_0\|^2}, \quad \text{(A.76)}\end{aligned}$$

where the second line first multiplies with $a^{-1}/a^{-1} = 1$ and subsequently with $\rho^2/\rho^2 = 1$.

To optimize the bound with respect to ρ, we consider the first-order condition:

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\rho}\frac{\sigma^2+\rho^2\sigma_\xi^2}{(\rho+\alpha)^2-\rho^2\|\boldsymbol{\Phi}_0\|^2} \\ &= \frac{2\rho\sigma_\xi^2((\rho+\alpha)^2-\rho^2\|\boldsymbol{\Phi}_0\|^2)-(\sigma^2+\rho^2\sigma_\xi^2)(2(\rho+\alpha)-2\rho\|\boldsymbol{\Phi}_0\|^2)}{((\rho+\alpha)^2-\rho^2\|\boldsymbol{\Phi}_0\|^2)^2} \\ &= \frac{2\rho\sigma_\xi^2((\rho+\alpha)^2-\rho^2\|\boldsymbol{\Phi}_0\|^2)-(\sigma^2+\rho^2\sigma_\xi^2)(2(\rho+\alpha)-2\rho\|\boldsymbol{\Phi}_0\|^2)}{\rho^4(a^{-1}+c)^2}, \quad \text{(A.77)}\end{aligned}$$

where the second line multiplies with $\frac{1}{\rho^4\rho^{-4}} = 1$ followed by using the definition of a and c. The contraction condition $ac < 1$ is assumed to hold, which implies that the denominator in (A.77) is positive; that is, $ac < 1 \Rightarrow 1-ac > 0 \Rightarrow a^{-1}-c > 0 \Rightarrow \rho^4(a^{-1}+c)^2 > 0$. This means that to solve the first-order condition, we need to set the numerator equal to 0, i.e.,

$$\begin{aligned}0 &= 2\rho\sigma_\xi^2((\rho+\alpha)^2-\rho^2\|\boldsymbol{\Phi}_0\|^2)-(\sigma^2+\rho^2\sigma_\xi^2)(2(\rho+\alpha)-2\rho\|\boldsymbol{\Phi}_0\|^2) \\ &= 2\rho\sigma_\xi^2(\rho+\alpha)^2-2\rho^3\sigma_\xi^2\|\boldsymbol{\Phi}_0\|^2-\sigma^2(2(\rho+\alpha)-2\rho\|\boldsymbol{\Phi}_0\|^2)-2\rho^2\sigma_\xi^2(\rho+\alpha)+2\rho^3\sigma_\xi^2\|\boldsymbol{\Phi}_0\|^2 \\ &= 2\rho\sigma_\xi^2(\rho+\alpha)^2-2\sigma^2(\rho+\alpha)+2\rho\sigma^2\|\boldsymbol{\Phi}_0\|^2-2\rho^2\sigma_\xi^2(\rho+\alpha) \\ &= 2\rho^3\sigma_\xi^2+2\rho\sigma_\xi^2\alpha^2+4\rho^2\sigma_\xi^2\alpha-2\rho\sigma^2-2\sigma^2\alpha+2\rho\sigma^2\|\boldsymbol{\Phi}_0\|^2-2\rho^3\sigma_\xi^2-2\rho^2\sigma_\xi^2\alpha \\ &= 2\rho\sigma_\xi^2\alpha^2+4\rho^2\sigma_\xi^2\alpha-2\rho\sigma^2-2\sigma^2\alpha+2\rho\sigma^2\|\boldsymbol{\Phi}_0\|^2-2\rho^2\sigma_\xi^2\alpha \\ &= \rho^2(2\sigma_\xi^2\alpha)+\rho(2\sigma_\xi^2\alpha^2-2\sigma^2(1-\|\boldsymbol{\Phi}_0\|^2))-2\sigma^2\alpha \\ &= A\rho^2+B\rho+C, \quad \text{(A.78)}\end{aligned}$$

which yields a quadratic equation in ρ with coefficients $A = 2\sigma_\xi^2\alpha$, $B = 2\sigma_\xi^2\alpha^2 - 2\sigma^2(1-\|\boldsymbol{\Phi}_0\|^2)$ and $C = -2\sigma^2\alpha$ and solution

$$\rho_\pm = \frac{-B \pm \sqrt{B^2-4AC}}{2A}. \quad \text{(A.79)}$$

Because $AC = -4\sigma_\xi^2\sigma^2\alpha^2 < 0$ and $A = 2\sigma_\xi^2\alpha > 0$, we have that $\rho_- < 0$ and $\rho_+ > 0$, the latter is therefore the solution to the minimization problem at hand. Specifically, the final

result reads

$$\rho_\star = \rho_+ = \frac{-2\sigma_\xi^2\alpha^2 + 2\sigma^2(1-\|\mathbf{\Phi}_0\|^2) + \sqrt{4(\sigma_\xi^2\alpha^2 - \sigma^2(1-\|\mathbf{\Phi}_0\|^2)^2 + 16\sigma_\xi^2\sigma^2\alpha^2}}{4\sigma_\xi^2\alpha}$$
$$= \frac{\sigma^2(1-\|\mathbf{\Phi}_0\|^2) - \alpha^2\sigma_\xi^2 + \sqrt{(\alpha^2\sigma_\xi^2 - \sigma^2(1-\|\mathbf{\Phi}_0\|^2)^2 + 4\alpha^2\sigma_\xi^2\sigma^2}}{2\alpha\sigma_\xi^2}, \tag{A.80}$$

where the second line multiplies by $(1/2)/(1/2) = 1$ and rearranges. Because $A = 2\sigma_\xi^2\alpha > 0$, we have that the numerator of (A.77) is a convex parabola with largest root $\rho_\star$, which means that it becomes negative for $\rho \in (0, \rho_\star)$ and positive for $\rho \in (\rho_\star, \infty)$. Combined with the fact that the denominator of (A.77) is always positive, we have that the bound is decreasing in ρ on the interval $(0, \rho_\star)$ and increasing on the interval $(\rho_\star, \infty)$. In sum, this means that the stationary point $\rho_\star$ is indeed the global minimum.

A.8 Example 3

For clarity, we omit all superscripts im. Under Assumptions 1–5, the asymptotic (Euclidean) MSE bound for the ISD filter is obtained by substituting $a = \left(\frac{\rho}{\rho+\alpha}\right)^2$, $b = a\rho^{-1}\sigma^2$, $c = \|\mathbf{\Phi_0}\|_2^2$, and $d = \rho\sigma_\xi^2$ in Equation (19), and multiplying by η to obtain

$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\sigma^2 + \sigma_\xi^2\rho^2}{\alpha^2 + 2\alpha\rho + (1-\|\mathbf{\Phi_0}\|_2^2)\rho^2} = \frac{\sigma\eta^2 + \sigma_\xi^2}{\alpha^2\eta^2 + 2\alpha\eta + (1-\|\mathbf{\Phi_0}\|_2^2)}, \tag{A.81}$$

where $\eta := 1/\rho$. In the context of Example 3, $\mathbf{\Phi_0} = \mathbf{\Phi}^\star = \phi_0$, $\alpha = -\nabla^2(y_t|\theta_{t|t-1}) = 1/\sigma_\varepsilon^2$ and $\sigma^2 = \mathbb{E}_{y_t}(\|\nabla(y_t|\vartheta_t)\|^2) = 1/\sigma_\varepsilon^2$. The postulated log-likelihood function is

$$\log p(y_t|\theta_{t|t-1}) = -\frac{1}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{(y_t - \theta_{t|t-1})^2}{2\sigma_\varepsilon^2}. \tag{A.82}$$

Taking the first derivative with respect to $\theta_{t|t-1}$

$$\nabla \log p(y_t|\theta_{t|t-1}) = \frac{y_t - \theta_{t|t-1}}{\sigma_\varepsilon^2}. \tag{A.83}$$

The second derivative with respect to $\theta_{t|t-1}$ is

$$\nabla^2 \log p(y_t|\theta_{t|t-1}) = -\frac{1}{\sigma_\varepsilon^2}. \tag{A.84}$$

Thus, $\alpha = 1/\sigma_\varepsilon^2$. Now, for $y_t \sim \mathrm{N}(\vartheta_t, \sigma_\varepsilon^2)$, the gradient is

$$\nabla \log p(y_t|\vartheta_t) = \frac{y_t - \vartheta_t}{\sigma_\varepsilon^2}, \tag{A.85}$$

and taking the squared norm, we obtain

$$\|\nabla \log p(y_t|\vartheta_t)\|^2 = \frac{(y_t - \vartheta_t)^2}{\sigma_\varepsilon^4}. \tag{A.86}$$

Taking the expectation with respect to y_t, we get:

$$\mathop{\mathbb{E}}_{\boldsymbol{y}_t}\left(\|\nabla \log p(y_t|\vartheta_t)\|^2\right) = \frac{\mathop{\mathbb{E}}_{\boldsymbol{y}_t}[(y_t - \vartheta_t)^2]}{\sigma_\varepsilon^4} = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^4} = \frac{1}{\sigma_\varepsilon^2}. \tag{A.87}$$

Thus, $\sigma^2 = \frac{1}{\sigma_\varepsilon^2}$. Next, we substitute $\boldsymbol{\Phi_0} = \phi_0$, $\alpha = 1/\sigma_\varepsilon^2$ and $\sigma^2 = 1/\sigma_\varepsilon^2$ in equation (A.81), and subsequently multiply the numerator and denominator by $\sigma_\epsilon^4 > 0$, to obtain

$$\limsup_{t\to\infty} \mathrm{MSE}_{t|t} \leq \frac{\sigma_\varepsilon^2\eta^2 + \sigma_\xi^2\sigma_\varepsilon^4}{\eta^2 + 2\sigma_\varepsilon^2\eta + (1-\phi_0^2)\sigma_\varepsilon^4}. \tag{A.88}$$

To find the (optimal) learning rate $\eta > 0$ minimizing the upper bound, we take the derivative with respect to η and equate to zero

$$\frac{[2\sigma_\varepsilon^2\eta]\left[2\sigma_\varepsilon^2\eta + \eta^2 + (1-\phi_0^2)\,\sigma_\varepsilon^4\right] - \left[\sigma_\varepsilon^2\eta^2 + \sigma_\xi^2\sigma_\varepsilon^4\right]\left[2\sigma_\varepsilon^2 + 2\eta\right]}{\left(\eta^2 + 2\sigma_\varepsilon^2\eta + (1-\phi_0^2)\sigma_\varepsilon^4\right)^2} = 0. \tag{A.89}$$

If $\eta^2 + 2\sigma_\varepsilon^2\eta + (1-\phi_0^2)\sigma_\varepsilon^4 \neq 0$, this is equivalent to

$$4\sigma_\varepsilon^4\eta^2 + 2\sigma_\varepsilon^2\eta^3 + 2\sigma_\varepsilon^6\eta\left(1-\phi_0^2\right) - \left(2\sigma_\varepsilon^4\eta^2 + 2\sigma_\varepsilon^2\eta^3 + 2\sigma_\xi^2\sigma_\varepsilon^6 + 2\sigma_\xi^2\sigma_\varepsilon^4\eta\right) = 0. \tag{A.90}$$

Canceling out $2\sigma_\varepsilon^2\eta^3$, using that $4\sigma_\varepsilon^4\eta^2 - 2\sigma_\varepsilon^4\eta^2 = 2\sigma_\varepsilon^4\eta^2$, and that $2\sigma_\varepsilon^6\eta(1-\phi_0^2) - 2\sigma_\xi^2\sigma_\varepsilon^4\eta = \eta\left(2\sigma_\varepsilon^6\eta(1-\phi_0^2) - 2\sigma_\xi^2\sigma_\varepsilon^4\right)$, combining the terms we get a quadratic equation in η

$$2\sigma_\varepsilon^4\eta^2 + \eta\left(2\sigma_\varepsilon^6\left(1-\phi_0^2\right) - 2\sigma_\xi^2\sigma_\varepsilon^4\right) - 2\sigma_\xi^2\sigma_\varepsilon^6 = 0. \tag{A.91}$$

Dividing each term by $2\sigma_\varepsilon^4 > 0$ leads to

$$\eta^2 + \eta\left(\sigma_\varepsilon^2\left(1-\phi_0^2\right) - \sigma_\xi^2\right) - \sigma_\xi^2\sigma_\varepsilon^2 = 0. \tag{A.92}$$

Solving for $\eta_\star$ using the quadratic formula, we obtain

$$\eta_\pm = \frac{\sigma_\xi^2 - \sigma_\varepsilon^2\left(1-\phi_0^2\right) \pm \sqrt{D}}{2}, \tag{A.93}$$

where $D = (\sigma_\varepsilon^2(1-\phi_0^2) - \sigma_\xi^2)^2 + 4\sigma_\xi^2\sigma_\varepsilon^2 = \sigma_\xi^4 + \sigma_\varepsilon^4(1-\phi_0^2)^2 + 2\sigma_\xi^2\sigma_\varepsilon^2(1+\phi_0^2)$. Because $\sqrt{D} > \sigma_\xi^2 - \sigma_\varepsilon^2\left(1-\phi_0^2\right)$, we have $\eta_+ > 0$ and $\eta_- < 0$. Only the positive solution is of interest. Specifically, we have

$$\eta_\star = \eta_+ = \frac{\sigma_\xi^2 - \sigma_\varepsilon^2\left(1-\phi_0^2\right) + \sqrt{\sigma_\xi^4 + \sigma_\varepsilon^4\left(1-\phi_0^2\right)^2 + 2\sigma_\xi^2\sigma_\varepsilon^2\left(1+\phi_0^2\right)}}{2}. \tag{A.94}$$

In the case of a local level model $\phi_0 = 1$, and the optimal learning rate reduces to

$$\eta_\star = \frac{\sigma_\xi^2 + \sqrt{\sigma_\xi^4 + 4\sigma_\xi^2\sigma_\varepsilon^2}}{2} = \frac{\sigma_\varepsilon^2}{2}\left(\frac{\sigma_\xi^2}{\sigma_\varepsilon^2} + \sqrt{\frac{\sigma_\xi^4}{\sigma_\varepsilon^4} + 4\frac{\sigma_\xi^2}{\sigma_\varepsilon^2}}\right), \tag{A.95}$$

which is exactly the steady state Kalman filter covariance of the prediction for the local level model as given in Durbin and Koopman (2012, p. 37) or Harvey (1989, p. 175), where $\sigma_\xi^2/\sigma_\varepsilon^2$ is the signal-to-noise ratio.

B Further results

B.1 Detailed discussion of ISD and ESD updates (4)–(5)

In optimization (4), 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

$$\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - \frac{1}{2}\|\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\|_{\boldsymbol{P}_t}^2 \;\geq\; \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{im}}),$$

where the left-hand side is the optimized value. After rearrangement, this implies

$$\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{im}}) \;\geq\; 1/2\|\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\|_{\boldsymbol{P}_t}^2 \;\geq\; 0,$$

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 $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{y}_t|\boldsymbol{\theta})$ 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.

In contrast, the solution (2) 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 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 mechanism $\boldsymbol{H}_t\nabla\ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t-1}^{\text{ex}})$ 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.

B.2 Detailed comparison of (7) with score-driven models

In equation (7), we could take $\boldsymbol{H}_t = \boldsymbol{H}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_{t|t-1}^{\text{ex}})^{-\zeta}$ for an arbitrary $\boldsymbol{H} \in \mathbb{R}^{k\times k}$ (thereby potentially violating our usual requirement $\boldsymbol{H}_t \succ \mathbf{O}_k$) and then re-parametrize by defining

$\boldsymbol{A} := \boldsymbol{\Phi}\boldsymbol{H}\mathbb{R}^{k\times k}$. This procedure would allow us to recover standard score-driven models, which typically write the matrix product $\boldsymbol{A}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\text{ex}}_{t|t-1})^{-\zeta}$ in front of the score (e.g. Creal et al., 2013, eq. 4-5). The downside, however, is that the resulting filter may fail to select a direction of ascent. Here we argue that maintaining $\boldsymbol{H}_t \succ \mathbf{O}_k$ is paramount. For example, the symmetrized sum $\boldsymbol{H}_t = (\boldsymbol{H}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\text{ex}}_{t|t-1})^{-\zeta} + \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\text{ex}}_{t|t-1})^{-\zeta}\boldsymbol{H})/2$ with a static $\boldsymbol{H} \succ \mathbf{O}_k$ would be permissible. This particular choice would cause the matrix in front of the score to have a special structure parametrized in terms of $\boldsymbol{H} \succ \mathbf{O}_k$, which contains fewer free parameters than the unrestricted $\boldsymbol{A} \in \mathbb{R}^{k\times k}$ as in Creal et al. (2013, eq. 4-5).

While score-driven filters in the literature can be obtained from (7) by relaxing some restrictions on our learning-rate matrix $\boldsymbol{H}_t$, some important differences remain. **First**, this literature has not considered splitting up the prediction-to-prediction recursion into distinct prediction and updating steps. Differentiating between both steps is conceptually useful, as new information is revealed during the update (but not the prediction) step, and practically useful, as it may be beneficial to distinguish between predictions and nowcasts. **Second**, ISD updates (4) enable the incorporation of parameter constraints without necessitating (but still allowing) link functions. **Third**, the ESD prediction-to-prediction recursion (7) illustrates that, in order to select a direction of ascent, the matrix in front of the score should have a special structure—an observation that, in the literature on score-driven models, has somehow escaped attention. **Fourth**, the literature on score-driven models has stopped short of recognizing that the ESD update (2) is, in fact, the solution of the linearized optimization problem (5). The lack of this connection being made has prohibited researchers from considering higher-order approximations (e.g., quadratic rather than linear) or the "full" (i.e., non-approximated) version (4). **Fifth**, the majority of this literature assumes that the (explicit) score-driven filters are correctly specified in the sense that the DGP is also score driven. Here, we take the more pessimistic view that the DGP remains largely unknown, while we can only hope to show that the (misspecified) ISD and ESD filters are stable over time (see Section 3), while tracking the (pseudo-)true time-varying parameter reasonably accurately (see Section 4).

B.3 Details for Section 5.1

Figure B.1 presents the MSE performance of the ISD and ESD filters alongside three recent competitors in the setting of Section 5.1. In this case, however, we follow Cutler et al. (2023) by setting $\alpha = \beta = 1$. As a result, the "momentum" term in the ONM algorithm vanishes, reducing it to the stochastic gradient descent method from Madden et al. (2021). As opposed to when $\alpha = 1, \beta = 40$, the learning rate in the Cutler et al. (2023) algorithm no

longer remains fixed at its cap, $1/(2/\beta)$. While the Cutler et al. (2023) algorithm performs well for small t, its empirical performance is quickly surpassed by both the ISD and ESD filters, consistent with their stronger asymptotic performance guarantees.

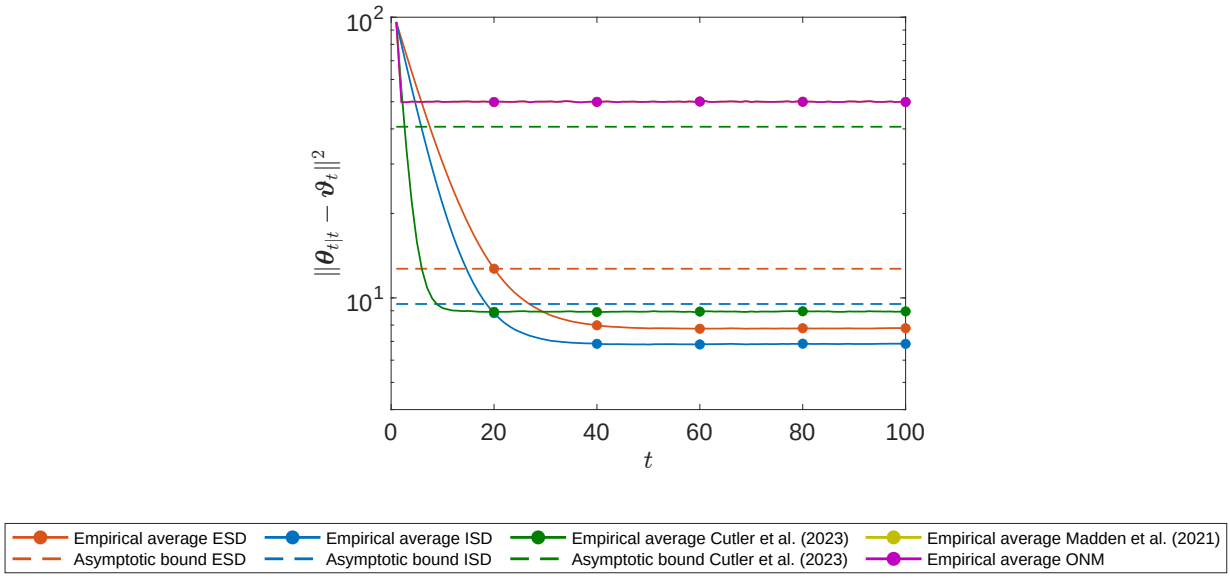

Figure B.1: Semilog plot of guaranteed bounds and empirical tracking errors for least-squares recovery with respect to iteration t. Empirical averages are computed over 10,000 trials. Default parameter values: $\alpha = 1, \beta = 1, \sigma = 10, \sigma_\xi = 1, \eta = \eta_\star$, $k = 50$, $n = 100$.

B.4 Details for Section 5.2

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

DGP Type	Distribution	**Link function**	**Density** $p(y_t\|\theta_t)$	**Score** $\nabla\ell(y_t\|\theta_t)$	**Negative Hessian** $-\nabla^2\ell(y_t\|\theta_t)$	**Information**
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.	$\lambda_t = \exp(\theta_t)$	$\frac{\Gamma(\kappa+y_t)}{\Gamma(\kappa)\Gamma(y_t+1)}\left(\frac{\kappa}{\kappa+\lambda_t}\right)^{\kappa}\left(\frac{\lambda_t}{\kappa+\lambda_t}\right)^{y_t}$	$y_t - \frac{\lambda_t(\kappa+y_t)}{\kappa+\lambda_t}$	$\frac{\kappa\lambda_t(\kappa+y_t)}{(\kappa+\lambda_t)^2}$	$\frac{\kappa\,\lambda_t}{\kappa+\lambda_t}$
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	$\beta_t = \exp(\theta_t)$	$\frac{y_t^{\kappa-1}\exp(-y_t/\beta_t)}{\Gamma(\kappa)\beta_t^{\kappa}}$	$\frac{y_t}{\beta_t} - \kappa$	$\frac{y_t}{\beta_t}$	κ
Duration	Weibull	$\beta_t = \exp(\theta_t)$	$\frac{\kappa\ (y_t/\beta_t)^{\kappa-1}}{\beta_t\exp\{(y_t/\beta_t)^{\kappa}\}}$	$\kappa\left(\frac{y_t}{\beta_t}\right)^{\kappa} - \kappa$	$\kappa^2\left(\frac{y_t}{\beta_t}\right)^{\kappa}$	κ^2
Volatility	Gaussian	$\sigma_t^2 = \exp(\theta_t)$	$\frac{\exp\{-y_t^2/(2\sigma_t^2)\}}{\{2\pi\sigma_t^2\}^{1/2}}$	$\frac{y_t^2}{2\sigma_t^2} - \frac{1}{2}$	$\frac{y_t^2}{2\sigma_t^2}$	$\frac{1}{2}$
Volatility	Student's t	$\sigma_t^2 = \exp(\theta_t)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)\left(1+\frac{y_t^2}{(\nu-2)\sigma_t^2}\right)^{-\frac{\nu+1}{2}}}{\sqrt{(\nu-2)\pi}\,\Gamma\left(\nu/2\right)\sigma_t}$	$\frac{\omega_t\, y_t^2}{2\sigma_t^2} - \frac{1}{2}$ $\omega_t := \frac{\nu+1}{\nu-2+y_t^2/\sigma_t^2}$	$\frac{\nu-2}{\nu+1}\,\frac{\omega_t^2\, y_t^2}{2\sigma_t^2}$	$\frac{\nu}{2\nu+6}$
Dependence	Gaussian	$\rho_t = \frac{1-\exp(-\theta_t)}{1+\exp(-\theta_t)}$	$\frac{\exp\left\{-\frac{y_{1t}^2+y_{2t}^2-2\rho_t y_{1t} y_{2t}}{2(1-\rho_t^2)}\right\}}{2\pi\sqrt{1-\rho_t^2}}$	$\frac{\rho_t}{2} + \frac{1}{2}\frac{z_{1t}\, z_{2t}}{1-\rho_t^2}$ $z_{1t} := y_{1t} - \rho_t y_{2t}$ $z_{2t} := y_{2t} - \rho_t y_{1t}$	$0 \not\leq \frac{1}{4}\frac{z_{1t}^2+z_{2t}^2}{1-\rho_t^2} - \frac{1-\rho_t^2}{4}$	$\frac{1+\rho_t^2}{4}$
Dependence	Student's t	$\rho_t = \frac{1-\exp(-\theta_t)}{1+\exp(-\theta_t)}$	$\frac{\nu\left(1+\frac{y_{1t}^2+y_{2t}^2-2\rho_t y_{1t} y_{2t}}{(\nu-2)(1-\rho_t^2)}\right)^{-\frac{\nu+2}{2}}}{2\pi(\nu-2)\sqrt{1-\rho_t^2}}$	$\frac{\rho_t}{2} + \frac{\omega_t}{2}\frac{z_{1t}\, z_{2t}}{1-\rho_t^2}$ $z_{1t} := y_{1t} - \rho_t y_{2t}$ $z_{2t} := y_{2t} - \rho_t y_{1t}$	$0 \not\leq \frac{\omega_t}{4}\frac{z_{1t}^2+z_{2t}^2}{1-\rho_t^2} - \frac{1-\rho_t^2}{4} - \frac{1}{2}\frac{\omega_t^2}{\nu+2}\frac{z_{1t}^2\, z_{2t}^2}{(1-\rho_t^2)^2}$ $\omega_t := \frac{\nu+2}{\nu-2+\frac{y_{1t}^2+y_{2t}^2-2\rho_t y_{1t} y_{2t}}{1-\rho_t^2}}$	$\frac{2+\nu(1+\rho_t^2)}{4(\nu+4)}$

Note: The table contains nine distributions and link functions taken from Koopman et al. (2016). To aid the comparison (but at the expense of some consistency), we retain most of their parameter notation, even though some symbols are also used in the main text for different quantities. We assume that the observation density is known, i.e. $p^0(\cdot|\vartheta_t) = p(\cdot|\theta_t)$ (hence, $\vartheta_t = \theta_t$). In each case, the state-transition equation is $\vartheta_t = (1-\phi_0)\omega + \phi_0\vartheta_{t-1} + \xi_t$ where $\xi_t \sim$ i.i.d.$(0, \sigma_\xi^2)$ being distributed either as a Gaussian or Student's t variate.

Table B.2: Static (hyper-)parameters for DGPs in Table B.1 in the low-volatility setting

Type	**Distribution**	ω_0	ϕ_0	σ_ξ	**Shape**
Count	Poisson	0.00	0.97	0.15	
Count	Negative binomial	0.00	0.97	0.15	$\kappa = 4$
Intensity	Exponential	0.00	0.97	0.15	
Duration	Gamma	0.00	0.97	0.15	$\kappa = 1.5$
Duration	Weibull	0.00	0.97	0.15	$\kappa = 1.2$
Volatility	Gaussian	0.00	0.97	0.15	
Volatility	Student's t	0.00	0.97	0.15	$\nu = 6$
Dependence	Gaussian	0.00	0.97	0.15	
Dependence	Student's t	0.00	0.97	0.15	$\nu = 6$

Table B.3: Out-of-sample MSE of $\{\theta^j_{t|t-1}\}$ for $j \in \{\text{im}, \text{ex}\}$ relative to true states $\{\vartheta_t\}$ under Gaussian state increments.

DGP		**Assum. 1**		**MSE bound?**		**Low vol.** ($\sigma_\xi = 0.15$)		**Med. vol** ($\sigma_\xi = 0.30$)		**High vol.** ($\sigma_\xi = 0.60$)	
Type	Distribution	α	β	ISD	ESD	ISD	ESD	ISD	ESD	ISD	ESD
Count	Poisson	0	∞	✓	✗	0.146	0.148	0.405	∞	1.641	∞
Count	Neg. Binomial	0	∞	✓	✗	0.158	0.160	0.430	0.452	1.644	1.853
Intensity	Exponential	0	∞	✓	✗	0.146	0.150	0.368	0.423	1.028	2.243
Duration	Gamma	0	∞	✓	✗	0.157	0.163	0.473	0.546	1.285	2.237
Duration	Weibull	0	∞	✓	✗	0.125	0.128	0.306	0.329	0.791	0.962
Volatility	Gaussian	0	∞	✓	✗	0.192	0.198	0.503	0.609	1.463	4.684
Volatility	Student's t	0	$\frac{\nu+1}{8}$	✓	✓	0.226	0.227	0.610	0.617	1.550	1.601
Dependence	Gaussian	$-\frac{1}{4}$	∞	✓	✗	0.239	0.231†	0.596	∞	1.561	∞
Dependence	Student's t	$-\frac{1}{4}$	$\frac{\nu+1}{4}$	✓	✓	0.251	0.252	0.620	0.625	1.499	1.529

Note: MSE = mean squared error. ISD = implicit score driven. ESD = explicit score driven. † = For the Gaussian dependence model, the ESD filtered path diverged for a single (out-of-sample) path in the low-volatility setting. For simplicity, we ignore this path and report a finite MSE.

B.5 Details for Section 5.3

Here, we provide additional implementation details for the ESD and ISD filters, as well as the MSE bounds for the latter from Section 5.3.

ESD filter. For the ESD filters with *exponential* link functions, $\mu_{t|t}^{\text{ex}} = \exp(\theta_{t|t}^{\text{ex}})$, we define the learning rates as $\boldsymbol{H}_t^{\text{ex}} = \eta^{\text{ex}} \mathcal{I}(\theta_{t|t-1}^{\text{ex}})^{-\zeta}$, where $\mathcal{I}(\theta) = \exp(\theta)$ represents the Fisher information and $\zeta \in \{0, 1/2, 1\}$ determines the scaling of the score function. The ESD update becomes

$$\theta_{t|t}^{\text{ex}} = \theta_{t|t-1}^{\text{ex}} + \eta^{\text{ex}} \exp(-\zeta\theta_{t|t-1}^{\text{ex}})(y_t - \exp(\theta_{t|t-1}^{\text{ex}})). \tag{B.96}$$

Combined with the prediction step (3), this leads to the standard score-driven prediction-to-prediction recursion (e.g. Creal et al., 2013, p. 779).

ISD filter. The ISD filter with *exponential* link function, $\mu_{t|t}^{\text{im}} = \exp(\theta_{t|t}^{\text{im}})$, and static learning rate, $\boldsymbol{H}_t^{\text{im}} = \eta^{\text{im}}$ for all t, has update

$$\theta_{t|t}^{\text{im}} = \theta_{t|t-1}^{\text{im}} + \eta^{\text{im}}(y_t - \exp(\theta_{t|t}^{\text{im}})), \tag{B.97}$$

which is solved numerically using a standard Newton-Raphson algorithm (see Appendix B.6). The ISD filter with *quadratic* link function, $\mu_{t|t}^{\text{im}} = \theta_{t|t}^{\text{im}2}$, and static learning rate, has an implicit update that can be solved analytically as

$$\begin{aligned}\theta_{t|t}^{\text{im}} &= \underset{\theta\in\Theta}{\text{argmax}} \left\{ \ell\left(y_t \mid \theta\right) - \frac{1}{2\eta^{\text{im}}} \left(\theta - \theta_{t|t-1}^{\text{im}}\right)^2 \right\}, \\ &= \underset{\theta\in\Theta}{\text{argmax}} \left\{ y_t \log(\theta^2) - \theta^2 - \log(y_t!) - \frac{1}{2\eta^{\text{im}}} \left(\theta - \theta_{t|t-1}^{\text{im}}\right)^2 \right\}. \end{aligned} \tag{B.98}$$

Taking first-order conditions with respect to θ, and evaluating in $\theta = \theta_{t|t}^{\text{im}}$, yields

$$\frac{2y_t}{\theta_{t|t}^{\text{im}}} - 2\theta_{t|t}^{\text{im}} - \frac{1}{\eta^{\text{im}}} \left(\theta_{t|t}^{\text{im}} - \theta_{t|t-1}^{\text{im}}\right)^2, \tag{B.99}$$

where we assume $\theta_{t|t}^{\text{im}} \neq 0$. As $\eta^{\text{im}} > 0$, we multiply both sides by $\eta^{\text{im}}\theta_{t|t}^{\text{im}}$ to obtain

$$-(1 + 2\eta^{\text{im}})\theta_{t|t}^{\text{im}2} + \theta_{t|t-1}^{\text{im}}\theta_{t|t}^{\text{im}} + 2\eta^{\text{im}} y_t = 0. \tag{B.100}$$

Solving for $\theta_{t|t}^{\text{im}}$, yields

$$\theta_{t|t}^{\text{im}} = \frac{-\theta_{t|t-1}^{\text{im}} \pm \sqrt{\theta_{t|t-1}^2 + 8(1 + 2\eta^{\text{im}})\eta^{\text{im}} y_t}}{-2(1 + 2\eta^{\text{im}})}. \tag{B.101}$$

Multiplying the numerator and denominator by -1 and taking only the positive solution, we obtain the ISD update

$$\theta_{t|t}^{\text{im}} = \frac{\theta_{t|t-1}^{\text{im}} + \sqrt{\theta_{t|t-1}^{\text{im}2} + 8\eta^{\text{im}}(1 + 2\eta^{\text{im}}) y_t}}{2(1 + 2\eta^{\text{im}})}. \tag{B.102}$$

MSE bounds. The gradient noise (in Assumption 4) can be computed as

$$\begin{aligned}
\sigma^2 &= \sup_t \mathbb{E}\left[\left(\frac{\mathrm{d}\ell(y_t|\theta_t)}{\mathrm{d}\theta_t}\right)^2 |_{\theta_t=\theta_t^\star}\right] \\
&= \sup_t \mathbb{E}\left[\left(\frac{2y_t}{\theta_t} - 2\theta_t\right)^2 |_{\theta_t=\theta_t^\star}\right] \\
&= \sup_t \mathbb{E}\left[\left(\frac{4y_t^2}{\theta_t^2} - 8y_t + 4\theta_t^2\right) |_{\theta_t=\theta_t^\star}\right] \\
&= \sup_t \mathbb{E}\left[\frac{4y_t^2}{\mu_t} - 8y_t + 4\mu_t\right] \\
&= \sup_t \left[\frac{4}{\mu_t}(\mu_t^2 + \mu_t) - 8\mu_t + 4\mu_t\right] \\
&= \sup_t \left[4\mu_t + 4 - 4\mu_t\right] \\
&= 4,
\end{aligned} \tag{B.103}$$

where in the first line we used that $\ell(y_t|\theta_t) = y_t \log(\mu_t) - \mu_t - \log(y_t!) = y_t \log(\theta_t^2) - \theta_t^2 - \log(y_t!)$, using $\mu_t = (2/\alpha)\theta_t^2 = \theta_t^2$ (as $\alpha = 2$), hence $\mathrm{d}\ell(y_t|\theta_t)/\mathrm{d}\theta_t = 2y_t\theta_t/\theta_t^2 - 2\theta_t = 2y_t/\theta_t - 2\theta_t$. In the fourth line we used that the pseudo-true state is identified as $\theta_t^\star = \sqrt{\mu_t} = \exp(\vartheta_t/2) \in (0, \infty)$ for all t, and thus $\theta_t^{\star 2} = \exp(\vartheta_t) = \mu_t$. In the fifth line, we used that $\mathbb{E}(y_t) = \mu_t$ and $\mathbb{E}(y_t^2) = \mu_t^2 + \mu_t$ for all t. The MSE bound of the ISD filter with quadratic link function additionally depends on $q^2 = \sup_t \mathbb{E}[\|\theta_t^\star - \theta_{t-1}^\star\|^2] < \infty$ and $s_\omega^2 = \sup_t \mathbb{E}[\|\theta_t^\star - \omega\|] < \infty$, which are assumed to be given. Here, we compute these quantities analytically assuming ω_0, ϕ_0 and σ_ξ are known

$$\begin{aligned}
q^2 &= \sup_t \mathbb{E}_{\vartheta_t}\left[\left(\theta_t^\star - \theta_{t-1}^\star\right)^2\right] \\
&= \sup_t \mathbb{E}_{\vartheta_t}\left[\left(\sqrt{\mu_t} - \sqrt{\mu_{t-1}}\right)^2\right] \\
&= \sup_t \mathbb{E}_{\vartheta_t}\left[\mu_t + \mu_{t-1} - 2\sqrt{\exp(\vartheta_t + \vartheta_{t-1})}\right] \\
&= 2\exp\left(\omega_0 + \frac{\sigma_\xi^2}{2(1-\phi_0^2)}\right) - 2\sup_t \mathbb{E}_{\vartheta_t}\left[\sqrt{\exp(\vartheta_t + \vartheta_{t-1})}\right] \\
&= 2\exp\left(\omega_0 + \frac{\sigma_\xi^2}{2(1-\phi_0^2)}\right) - 2\exp\left(\omega_0 + \frac{(1+\phi_0)\sigma_\xi^2}{4(1-\phi_0^2)}\right) \\
&= 2\exp\left(\omega_0 + \frac{\sigma_\xi^2}{2(1-\phi_0^2)}\right) - 2\exp\left(\omega_0 + \frac{\sigma_\xi^2}{4(1-\phi_0)}\right), \quad \phi_0 = 0.98
\end{aligned} \tag{B.104}$$

$$\begin{aligned}
s_\omega^2 &= \sup_t \mathbb{E}\left[\|\theta_t^\star - \omega\|\right] \\
&= \sup_t \mathbb{E}_{\vartheta_t}\left[\mu_t + \omega^2 - 2\omega\sqrt{\mu_t}\right] \\
&= \sup_t \mathbb{E}_{\vartheta_t}\left[\mu_t\right] + \omega^2 - 2\omega \sup_t \mathbb{E}_{\vartheta_t}\left[\sqrt{\mu_t}\right]
\end{aligned}$$

$$= \exp\left(\omega_0 + \frac{\sigma_\xi^2}{2(1-\phi_0^2)}\right) + \omega^2 - 2\omega \exp\left(\frac{\omega_0}{2} + \frac{\sigma_\xi^2}{8(1-\phi_0^2)}\right), \tag{B.105}$$

where we used that $\mathbb{E}_{\vartheta_t}(\exp(\vartheta_t)) = \mathbb{E}_{\vartheta_t}(\mu_t)) = \exp(\omega_0 + \sigma_\xi^2/(2(1-\phi_0^2)))$, $\mathbb{E}_{\vartheta_t}(\sqrt{\exp(\vartheta_t + \vartheta_{t-1})}) = \exp(\omega_0 + (1+\phi_0)\sigma_\xi^2/(4(1-\phi_0^2)))$, and $\mathbb{E}_{\vartheta_t}(\sqrt{\exp(\vartheta_t)}) = \exp(\omega_0/2 + \sigma_\xi^2/(8(1-\phi_0^2)))$.

Figure B.2 plots the maximum likelihood-estimated learning rates of the ISD and ESD filters. The learning rates of the ISD filters both increase monotonically with the state variation, which is intuitive, as more sensitivity of the filter is needed to track more volatile states. The learning rates of the ESD filters, on the other hand, peak before declining, likely as an attempt to prevent divergence of the filter when the true states are volatile.

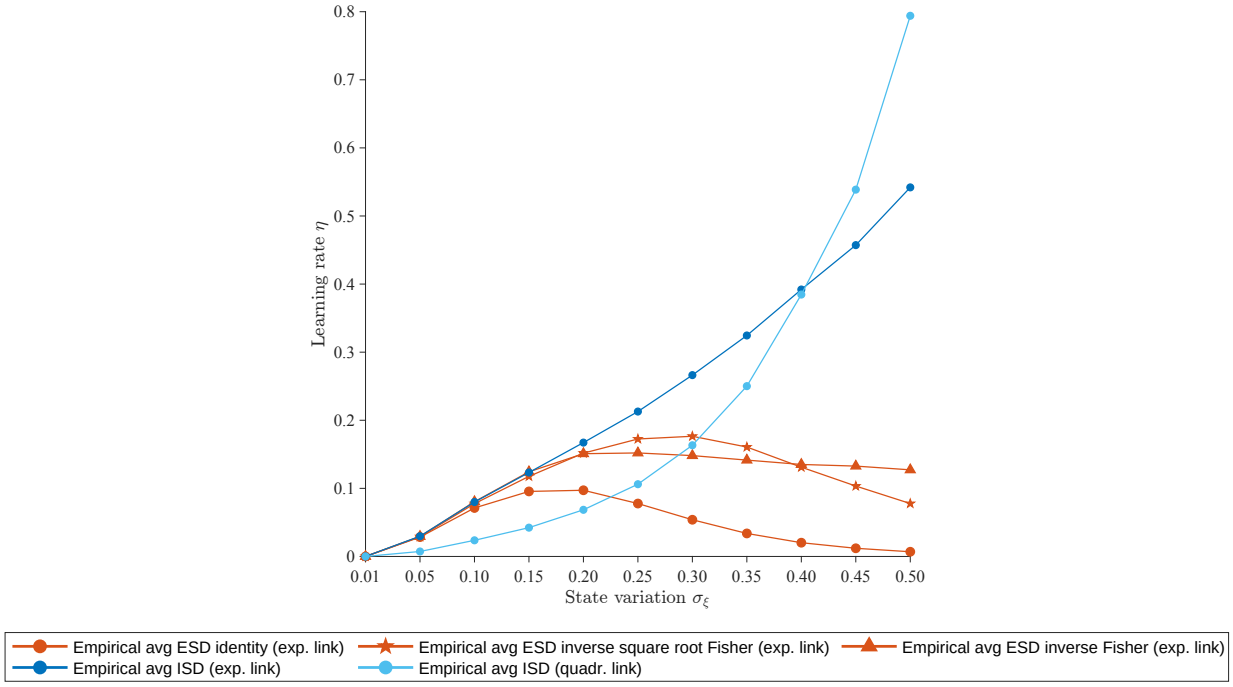

Figure B.2: Plot of the maximum likelihood-estimated learning rates for the ESD filter with identity scaling ($\zeta = 0$), inverse square root Fisher scaling ($\zeta = 1/2$), and inverse Fisher scaling ($\zeta = 1$), and the ISD filter, all using exponential link functions, along with the ISD filter using a quadratic link function.

B.6 Computing the implicit-gradient update

Standard Newton-Raphson iterates for solving the *implicit* update (1) read

$$\boldsymbol{\theta}_{t|t}^{\text{im}} \leftarrow \boldsymbol{\theta}_{t|t}^{\text{im}} + \left[\boldsymbol{P}_t - \nabla^2 \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}})\right]^{-1} \left[\nabla \ell(\boldsymbol{y}_t|\boldsymbol{\theta}_{t|t}^{\text{im}}) - \boldsymbol{P}_t\left(\boldsymbol{\theta}_{t|t}^{\text{im}} - \boldsymbol{\theta}_{t|t-1}^{\text{im}}\right)\right], \tag{B.106}$$

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 $\boldsymbol{\theta}_{t|t}^{\text{im}} \leftarrow \boldsymbol{\theta}_{t|t-1}^{\text{im}}$. 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 (e.g., Liu and Nocedal, 1989). When computational efficiency is critical, algorithm (B.106) 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 (2) 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 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 (B.106). In our simulation studies, $\sim$5 iterations typically provide an excellent level of accuracy.

B.7 Empirical results for level model with EGB2 innovations

Figure B.3 shows the results for the EGB2 distribution in Section 6. Figure B.3 is similar to Figure 4 in the main text, except the news impact curves for both filters are (continuous and) monotone increasing rather than re-descending.

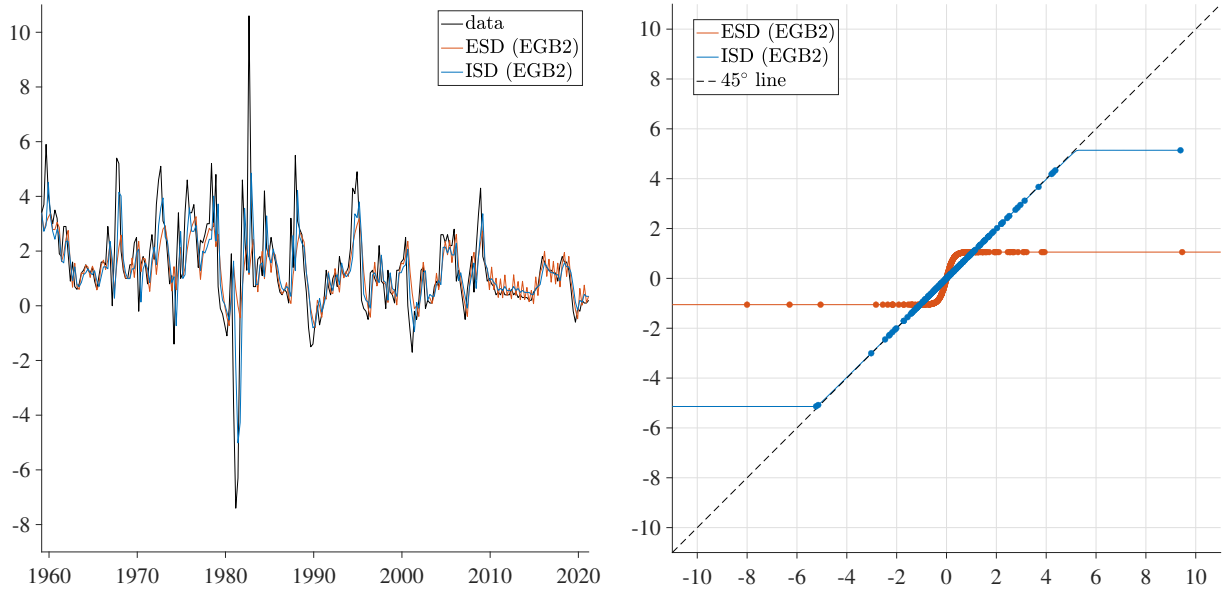

Figure B.3: Time-series data of T-bill rate spreads modeled using an EGB2 distribution with ESD and ISD predictions $\{\theta^j_{t|t-1}\}_t$ for $j \in \{\text{ex}, \text{im}\}$ (left) and associated impact curves (right), which shows $y_t - \theta_{t|t-1}$ (horizontally) versus $\theta_{t|t} - \theta_{t|t-1}$ (vertically).