TI 2022-066/III
Tinbergen Institute Discussion Paper

Implicit score-driven filters for time-varying parameter models

Revision: February 2025

Rutger-Jan Lange[1]
Bram van Os[2]
Dick van Dijk[3]

1 Erasmus University Rotterdam, Tinbergen Institute

2 Vrije Universiteit Amsterdam, Tinbergen Institute

3 Erasmus University Rotterdam, Tinbergen Institute

Implicit score-driven filters for time-varying parameter models*

Rutger-Jan Lange†, Bram van Os‡ and Dick van Dijk§

12 February 2025

Abstract

We propose an observation-driven modeling framework that permits time variation in the model parameters using an implicit score-driven (ISD) update. The ISD update maximizes the logarithmic observation density with respect to the parameter vector, while penalizing the weighted ℓ_2 norm relative to a one-step-ahead prediction. This yields an *implicit* stochastic-gradient update; we show that the popular class of *explicit* score-driven (ESD) models arises if the observation log density is linearly approximated around the prediction. By preserving the full density, the ISD update globalizes favorable local properties of the ESD update. Namely, for log-concave observation densities (even when misspecified), ISD filters are stable for any learning rate and globally contractive to a pseudo-truth. We demonstrate the usefulness of ISD filters in simulations and empirical illustrations in finance and macroeconomics.

Keywords: Implicit gradient, proximal-point method, stochastic-gradient descent, observation-driven models

*We thank the following individuals for their comments and suggestions: Christian Brownlees, Janneke van Brummelen, Leopoldo Catania, Guillaume Chevillon, Simon Donker van Heel, Gustavo Freire, Ana Galvao, Peter Hansen, Onno Kleen, Erik Kole, Andrew Patton and participants of the NESG 2022, CEF 2022, SoFiE 2022, Aarhus Econometrics Workshop 2022, UvA Econometrics Seminar 2022, SNDE Workshop for Young Scholars 2022, Barcelona Workshop in Financial Econometrics 2023, CREST 2023, and NBER-NSF 2023. Remaining errors are our own.

†Econometric Institute, Erasmus University Rotterdam (lange@ese.eur.nl)

‡Econometrics and Data Science Department, Vrije Universiteit Amsterdam (b.van.os2@vu.nl)

§Econometric Institute, Erasmus University Rotterdam (djvandijk@ese.eur.nl)

1 Introduction

Ample empirical evidence suggests it is often too restrictive to assume that model parameters remain constant for prolonged periods of time. In economics and finance, parameters are often found to be regime dependent or subject to structural breaks (e.g., Stock and Watson, 1996). Parameters may also change more gradually and not follow an easily discernible pattern, making it unclear how to update them after observing new data. In some cases, ex-post estimators can be constructed. For example, in ARCH-type models (see Teräsvirta, 2009, for an overview), time-varying volatility is updated using the squared shock, which provides an unbiased ex-post proxy of the true variance. In general, however, such proxies may be difficult to derive, inefficient, or nonexistent.

We propose a comprehensive framework that allows a model's parameters to be made time-varying in an observation-driven setting by means of an implicit score-driven (ISD) filter. Analogous to the Kalman (1960) filter, the proposed ISD filter alternates between a prediction step and, crucially, an update step. The ISD update step is the solution to an optimization problem that maximizes the log-likelihood contribution of the current observation subject to a weighted ℓ_2 penalty centered at the one-step-ahead prediction. The penalty weights are controlled by a positive-definite matrix, the inverse of which can be viewed as a learning-rate matrix. Optimizing the logarithmic likelihood allows new information to be efficiently incorporated, while the penalty term regularizes the extent to which the updated parameter deviates from its prediction. This ISD setup also enables automatic coordination of the updates of multiple interacting parameters and incorporation of constraints without necessitating parameter transformations.

The first-order condition corresponding to the ISD optimization problem can be formulated as an implicit stochastic-gradient update: implicit because the gradient is evaluated in the updated rather than the predicted parameter, and stochastic because it uses noisy data. In the optimization literature, such methods are known as proximal-point methods, and are recognized as inherently more stable than their explicit counterparts, which arise as first-order approximations. Moreover, implicit gradient approaches are guaranteed to improve the objective function—in our case, the log-likelihood contribution of the most recent observation. As we discuss in Section 1.1, explicit score-driven (ESD) updates do not share this desirable property, despite being widely used to track time-varying parameters.

The ISD filter has several attractive theoretical properties, as outlined below and demonstrated in detail throughout the paper. These properties are typically sought in observation-driven models, but rarely combined in a single framework. First, the ISD filter is invertible under mild and easily verifiable conditions; for example, concavity of the researcher-

postulated logarithmic density is, even when misspecified, typically sufficient (Theorem 1). Hence, any differences stemming from the initialization of the filter disappear almost surely and exponentially fast—a crucial property in the filtering literature (e.g., Bougerol, 1993, Straumann and Mikosch, 2006). Second, the ISD update is globally contractive in expectation to a small region around the pseudo-truth (Theorem 2). This means that, on average, the update is more accurate than the prediction on which it is based. This contraction is global in that the predictions may be arbitrarily poor; in fact, the largest improvements are expected for the worst predictions. Only when the prediction is very close to the (pseudo-)true parameter may the update be less accurate, as is unavoidable when using noisy data. The contraction property of the ISD filter is also robust in that it holds for an arbitrary (positive-definite) learning-rate matrix. This stands in contrast to ESD filters, which require additional stringent assumptions that limit the magnitude of the update; this is the price paid for relying on first-order approximations.

We demonstrate the theoretical and practical advantages of the ISD filter in simulation experiments and empirical illustrations. In simulations, we find that the ESD filter may diverge, even in relatively simple settings, while the ISD filter remains well-behaved. This finding appears to be new in the literature on parameter tracking (Section 1.1). In our empirical illustrations, we consider a linear regression of daily Microsoft equity returns on the market factor, where the regression coefficient (i.e., the slope) is made time-varying. Additionally, we consider growth-at-risk estimates captured by the lower quantiles of quarterly U.S. GDP growth. In this case, the ISD quantile update yields an implicit version of Engle and Manganelli's (2004) adaptive CAViaR model, with the advantage that the ISD update cannot be more extreme than the observation just received. This enhanced stability, together with simple parameter restrictions, ensures that jointly modeled quantiles remain properly ordered, thereby avoiding the common quantile-crossing problem.

In Section 2, we outline the ISD methodology and highlight the differences with conventional ESD models. Section 3 presents the theoretical properties, focusing on filter stability and optimality, while Section 4 discusses maximum-likelihood estimation of the static parameters. Sections 5 and 6 contain simulations and empirical illustrations, respectively. Finally, Section 7 concludes. All proofs are provided in the appendix.

1.1 Positioning in the literature

This paper intersects with two strands of literature that can be characterized along two axes, as visualized in Table 1, by (a) the stochastic-gradient method used (i.e., explicit or implicit) and (b) whether the focus is on "learning" or "tracking" (i.e., the parameter to be estimated

	Explicit gradient method	Implicit gradient method
Learning (static target)	SGD (e.g. Robbins and Monro, 1951; Amari, 1993; Bottou, 2012)	ISGD (e.g. Patrascu and Necoara, 2018; Asi and Duchi, 2019; Toulis et al., 2021)
Tracking (dynamic target)	ESD filter (e.g. Creal et al., 2013; Harvey, 2013; www.gasmodel.com)	**ISD filter** (this article)

Table 1: Overview of related methods.

Note: (I)SGD = (implicit) stochastic gradient descent. (I/E)SD = (implicit/explicit) score driven.

is static or dynamic). To the best of our knowledge, this paper is unique in the fourth quadrant, that is, in using an *implicit* gradient method to *track* a dynamic parameter.[1]

Regarding the first axis (implicit v. explicit gradient methods), the ISD filter is related to implicit gradient methods for static optimization problems, in particular Rockafellar's (1976) proximal-point algorithm, which combines a static target function to be optimized with a quadratic penalty involving some previous iterate. As our log-likelihood function involves (random) observations drawn from the true density, at every time step the ISD filter can be viewed as a stochastic proximal-point method (e.g., Bauschke et al., 2003; Ryu and Boyd, 2016; Bianchi, 2016; Patrascu and Necoara, 2018; Asi and Duchi, 2019). As is well known, the proximal optimization can be reformulated as an implicit stochastic-gradient step (e.g., Toulis and Airoldi, 2015; Toulis et al., 2016; Toulis and Airoldi, 2017; Toulis et al., 2021). Our approach is also related to online-learning methods that sequentially process data (e.g., Orabona, 2019; Cesa-Bianchi and Orabona, 2021), notably in machine-learning applications (e.g., Kulis and Bartlett, 2010). What differentiates our work from implicit gradient methods in the (stochastic) optimization literature is that we consider a setup in which the parameter to be estimated is dynamic.

Regarding the second axis (learning v. tracking), the ISD filter is related to the literature that uses explicit stochastic-gradient methods for tracking time-varying parameters. In particular, dynamic conditional score (DCS; Harvey, 2013) models and generalized autoregressive score (GAS; Creal et al., 2013) models use the (explicit) gradient of the log-likelihood function, known as the *score*, to update the time-varying parameters. This framework encompasses many established models, such as the GARCH model, and is popular for its ease of use and strong forecasting performance (e.g., Creal et al., 2014; Harvey and Luati, 2014; Koopman et al., 2016; Harvey and Lange, 2017; Opschoor et al., 2018; Gorgi, 2020). It has been used in ∼400 published articles; for a near-exhaustive list, see www.gasmodel.com. Recent survey articles (Artemova et al., 2022a; Artemova et al., 2022b; Harvey, 2022) have converged on the terminology of score-driven (SD) models. To align with this nomenclature

[1]The overview presented in Table 1 is not exhaustive; for example, we have left out all simulation-based approaches such as particle filters (e.g., Chopin and Papaspiliopoulos, 2020).

while highlighting the difference with our approach, we will refer to this (existing) model class as using *explicit* score-driven (ESD) filters. As this article demonstrates, ESD filters can be obtained within the ISD framework by replacing, at each time step, the logarithmic observation density by its local-linear approximation around the prediction—in the time-series literature, this insight is apparently new. We will show that avoiding the local-linear approximation has both theoretical and practical advantages.

2 Methodology

2.1 Implicit score-driven filters

We consider an $N \times 1$ variable of interest y_t, observed at times $t = 1, \dots, T$, drawn from a data-generating process (DGP) characterized by a density $p_0(\cdot|\theta_t^0, \psi^0, \mathcal{F}_{t-1})$. Hence, y_t is drawn at each time step from a conditional distribution, which is controlled by a $K_0 \times 1$ time-varying parameter vector θ_t^0 taking values in some parameter space Θ^0. Further, ψ^0 is a vector of static shape parameters, and $\mathcal{F}_{t-1}$ denotes the information set at time $t-1$, thus permitting dependence on exogenous variables and/or lags of y_t. For readability, the dependence on ψ^0 and $\mathcal{F}_{t-1}$ is suppressed; i.e., we write $p_0(\cdot|\theta_t^0)$ for $p_0(\cdot|\theta_t^0, \psi^0, \mathcal{F}_{t-1})$. The dynamics of the true process $\{\theta_t^0\}$ are left, for the most part, unspecified.

The aim of this paper is to devise a modeling framework that attempts to approximate the true observation density $p_0(\cdot|\theta_t^0)$. To this end, we consider filters that alternate between prediction and update steps. Specifically, let $p(\cdot|\theta_t)$ denote the researcher-postulated observation density, which may or may not be correctly specified, where θ_t denotes a $K \times 1$ vector of time-varying parameters that can take values in some non-empty convex parameter space $\Theta \subseteq \mathbb{R}^K$. As above, additional dependence on static shape parameters ψ and/or other information available at time $t-1$ is permitted, but suppressed for readability. We denote the predicted and updated parameter vectors by $\theta_{t|t-1} \in \Theta$ and $\theta_{t|t} \in \Theta$, which reflect the researcher's estimates of θ_t using the information available at times $t-1$ and t, respectively.

The main difficulty in working with time-varying parameter models lies in specifying how $\theta_{t|t}$ should be updated from $\theta_{t|t-1}$ after observing y_t. We argue that a sound update scheme should satisfy at least two criteria. First, the update should yield an improved fit of the observed data y_t in terms of the likelihood, i.e., $p(y_t|\theta_{t|t}) \geq p(y_t|\theta_{t|t-1})$. As we shall see, explicit score-driven filters generally fail to meet this requirement. Second, as each observation y_t is inherently noisy, it is desirable to regularize the extent to which the update $\theta_{t|t}$ deviates from the prediction $\theta_{t|t-1}$. Penalizing the magnitude of $\theta_{t|t} - \theta_{t|t-1}$ prohibits the filter from becoming excessively volatile.

To satisfy both criteria, we propose the class of implicit score-driven (ISD) filters. These filters perform the parameter update at time t by maximizing the researcher-postulated logarithmic observation density $\log p(y_t|\cdot)$ subject to a weighted ℓ_2 penalty centered at the prediction $\theta_{t|t-1}$. That is, the parameter update is defined as

$$\theta_{t|t} \;:=\; \underset{\theta\in\Theta}{\operatorname{argmax}}\; f(\theta|y_t, \theta_{t|t-1}, P_t), \tag{1}$$

where

$$f(\theta|y_t, \theta_{t|t-1}, P_t) \;:=\; \log p(y_t|\theta) \;-\; \frac{1}{2}\Big\|\theta - \theta_{t|t-1}\Big\|^2_{P_t}. \tag{2}$$

Here, $f(\theta|y_t, \theta_{t|t-1}, P_t)$ denotes the "regularized" log-likelihood contribution and $\|x\|^2_{P_t} = x'P_tx$ is the squared ℓ_2 norm with respect to a $K \times K$ positive-definite penalty matrix P_t. By formulating the parameter update as the solution to a maximization problem, the proposed method has several favorable characteristics. First, all information in the conditional density (as opposed to, e.g., moment information only) is utilized to update the parameter. Second, elements of the parameter update $\theta_{t|t}$ are automatically interdependent, because jointly they represent the solution to the multivariate optimization problem (1). Third, the update $\theta_{t|t}$ is automatically contained in the correct space Θ and does not require specification of a link function. In fact, we could constrain Θ to any non-empty convex subset, allowing for straightforward incorporation of a great variety of constraints. When θ_t contains positive-valued time-varying shape parameters (as in Section 5.3), optimization (1) automatically keeps them positive.

The ℓ_2 penalty yields tractable updates and can be interpreted as a second-order Taylor expansion around $\theta_{t|t-1}$ of a smooth loss function, where P_t acts as the Hessian. Furthermore, the ISD update defined in equations (1)–(2) takes a comparable form to Rockafellar's (1976) classic proximal-point algorithm, which similarly considers the optimization of a target function—in our case, the log-likelihood contribution of the (a priori random) realization y_t—subject to a quadratic penalty. For a fixed time step, therefore, the approach can be viewed as a stochastic proximal-point method (e.g., Bauschke et al., 2003; Asi and Duchi, 2019); the difference, as discussed in Section 1.1, is that we consider a moving target.

Update (1) can also be seen as computing the posterior mode in a (possibly misspecified) Bayesian framework, where the quadratic penalty corresponds to a Gaussian prior. This perspective highlights a link with Laplace approximation methods in both the Bayesian literature (e.g., Rue et al., 2009) and the frequentist literature (e.g., Koyama et al., 2010). Interestingly, (1) reduces to Kalman's level update if $p(\cdot|\theta)$ is a Gaussian density, while its mean is a linear transformation of θ, and the penalty matrix is taken to be the inverse of Kalman's predicted covariance matrix. Although the link between (least-squares) opti-

mization methods and the Kalman filter has been known at least since Bierman (1977) and Bertsekas (1996), it recently has attracted renewed interest in signal processing (e.g., Akyildiz et al., 2019), control theory (e.g., Simonetto and Massioni, 2024), and econometrics (e.g., Lange, 2024a; Lange, 2024b).

These numerous connections motivate the investigation of the proximal method (1) in a more *general* context, where the observation density may be non-Gaussian, θ may be unrelated to the mean, and the quadratic penalty does not necessarily correspond to a Gaussian prior. Rather, we attempt to remain, as far as possible, agnostic with regard to the true state dynamics. Specifically, we make no (hidden) assumption about the linearity or Gaussianity of the true states $\{\theta_t^0\}$. Consistent with this aim, update (1) is postulated as (part of) a *filter* or an *algorithm*—a conceptually distinct approach from imposing conditions on the DGP. We are interested in investigating the performance of the proposed algorithm, especially when some (or all) of the classic assumptions fail. Despite its simplicity and close connection with existing methods, the proposed method is—at this level of generality—new.

Whereas the DGP is unknown, the postulated density $p(\cdot|\theta)$ is under the researcher's control and thus, typically, it is analytically known. For this reason, the assumptions below relate only to this postulated density, and, as such, have the advantage of being practically verifiable. Assumptions 1 and 2 are standard in the optimization literature, ensuring the existence and uniqueness of the solution to the maximization problem (1). Assumptions 3 and 4 are convenient in allowing us to characterize its solution using a standard first-order condition. While this simplification is not strictly necessary (e.g., we could work with subgradients), it benefits the clarity of exposition and improves tractability.

Assumption 1 (Existence) *The solution set of* $\operatorname{argmax}_{\theta\in\Theta} f(\theta|y_t, \theta_{t|t-1}, P_t)$ *is non-empty with probability one.*

Assumption 2 (Strictly concave regularized log likelihood) $f(\theta|y_t, \theta_{t|t-1}, P_t)$ *is proper strictly concave in* θ, $\forall\theta \in \Theta$ *with probability one.*

Assumption 3 (Interior solution) $\theta_{t|t} \in \operatorname{Int}(\Theta)$ *with probability one.*

Assumption 4 (Differentiability) $\log p(y_t|\theta)$ *is at least (a) once or (b) twice continuously differentiable in* θ, $\forall\theta \in \operatorname{Int}(\Theta)$ *with probability one. When left unspecified, (b) holds.*

Assumptions 1 and 2 can typically be satisfied through a sufficiently large penalty P_t. Even if the postulated logarithmic density is badly behaved (e.g., non-concave or multimodal), as long as the penalty term is strong enough, update (1) remains well-behaved.

Under Assumptions 1 through 4a, the first-order condition for the parameter update $\theta_{t|t}$ in the maximization problem (1), i.e., $0 = \nabla(y_t|\theta_{t|t}) - P_t(\theta_{t|t} - \theta_{t|t-1})$, can be rearranged as

$$\theta_{t|t} \;=\; \theta_{t|t-1} \;+\; H_t\,\nabla(y_t|\theta_{t|t}), \tag{3}$$

where the inverse penalty $H_t := P_t^{-1}$ is referred to as the learning-rate matrix at time t and $\nabla(y_t|\theta_{t|t}) := (\partial \log p(y_t|\theta)/\partial\theta)|_{\theta=\theta_{t|t}}$ is the score evaluated in $\theta_{t|t}$. Representation (3) demonstrates that the ISD framework yields a gradient-type parameter update. The learning-rate matrix H_t controls the step size and allows for different learning rates and interactions between the different time-varying parameters. Crucially, the score is evaluated at the update $\theta_{t|t}$ rather than the prediction $\theta_{t|t-1}$. This means that update (3) is an *implicit* gradient method; i.e., the parameter update $\theta_{t|t}$ appears on both sides of the equation and is thus not immediately computable. Because the update $\theta_{t|t}$ is also stochastic—it is based on the a-priori random realization y_t—our framework is closely related to implicit stochastic-gradient methods (Section 1.1). While the first-order condition (3) may not allow a closed-form solution, Assumptions 2 and 4a guarantee that the global solution to optimization problem (1) can be found numerically using standard optimization techniques.

In the optimization literature, the learning-rate matrix H_t is often set to decrease over time (e.g., $H_t = \mathcal{O}(t^{-1})$), such that the parameter asymptotically converges to some constant pseudo-true value. We are interested in tracking a time-varying true parameter; hence, our filtered path must not converge over time, but remain responsive even asymptotically. To achieve this, we can keep H_t constant over time, i.e., set $H_t = H$ for all t, where H may contain static parameters to be estimated (Section 4).

To complete our filter setup, the ISD update step (1) is complemented with a prediction step that generates one-step-ahead forecasts. For simplicity, we consider a linear first-order specification as follows:

$$\theta_{t+1|t} \;=\; \omega \;+\; \Phi\,\theta_{t|t}, \tag{4}$$

where ω is a $K \times 1$ vector of constants and Φ is a $K \times K$ autoregressive matrix. Conditions ensuring stable recursions are discussed in the next section. The requirement $\theta_{t+1|t} \in \Theta$ can typically be fulfilled by appropriate parameter restrictions and/or link functions. Again we emphasize that the linear prediction step (4) is a property of the *filter*, not the DGP. Of course, the prediction step could be generalized to allow for non-linear and/or higher order dynamics if these were found to be relevant for a particular application. In economics and statistics, however, mean reversion is often important during the prediction step, when no additional information is available. In these cases, a more complicated structure may not yield immediate benefits.

2.2 Relationship with explicit score-driven filters

The ISD update (3) suggests a close connection with existing (i.e., explicit) score-driven models. Here we show that linearizing the logarithmic observation density in the ISD optimization problem (1) produces the familiar explicit gradient update. Specifically, suppose we approximate the logarithmic observation density in equation (2) using a first-order Taylor expansion around the prediction $\theta_{t|t-1}$, i.e., $\log p(y_t|\theta) \approx \log p(y_t|\theta_{t|t-1}) + \langle \theta - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1}) \rangle$, where $\langle x_1, x_2 \rangle := x_1' x_2$ denotes the inner product. To avoid boundary solutions, we consider maximization over the Euclidean space $\mathbb{R}^K$. Because the regularized log-likelihood contribution $f(\cdot|\cdot,\cdot,\cdot)$ in optimization problem (1) now becomes a linear target in combination with a quadratic penalty, the optimization can be performed in closed form. Indeed, the resulting linearized version of optimization (1) and associated first-order condition now read

$$\theta^{\text{ex}}_{t|t} := \underset{\theta \in \mathbb{R}^K}{\text{argmax}} \left\{ \log p(y_t|\theta_{t|t-1}) + \langle \theta - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1}) \rangle - \frac{1}{2} \|\theta - \theta_{t|t-1}\|^2_{P_t} \right\}, \tag{5}$$

$$\theta^{\text{ex}}_{t|t} \;=\; \theta_{t|t-1} \;+\; H_t \, \nabla(y_t|\theta_{t|t-1}), \tag{6}$$

where the explicit update is denoted $\theta^{\text{ex}}_{t|t}$ to differentiate it from the ISD update (3).

Combining the first-order condition (6) with the linear prediction step (4) reproduces the dynamic conditional score (DCS; Harvey, 2013) update or, equivalently, the generalized autoregressive score (GAS; Creal et al., 2013) update. These updates have collectively become known as *score driven* (Section 1.1); hence, we refer to equation (6) as the explicit score-driven (ESD) update. Combining the ESD update (6) with the linear prediction step (4), we obtain the prediction-to-prediction recursion that is standard in the ESD literature:

$$\theta^{\text{ex}}_{t+1|t} \;=\; \omega \;+\; A \, S_t \, \nabla(y_t|\theta_{t|t-1}) \;+\; \Phi \, \theta_{t|t-1}. \tag{7}$$

Here, we take a dynamic learning-rate matrix $H_t = H S_t$, consisting of a time-invariant learning-rate matrix H and a dynamic scaling matrix S_t known at time t, while $A := \Phi H$ is a combined coefficient matrix of static parameters. The scaling matrix S_t is often chosen based on the Fisher information of the postulated density (e.g., Artemova et al., 2022a). While the relationship between implicit and explicit updates is well-known in the optimization literature (e.g., Rockafellar, 1976), it is apparently novel in the time-series literature.

The inherent locality of the first-order Taylor expansion suggests that the ESD update may perform satisfactory only for "minor" updates. Notably, the ESD update does *not* ensure that the local fit is improved; that is, it is not guaranteed that $p(y_t|\theta^{\text{ex}}_{t|t}) \geq p(y_t|\theta_{t|t-1})$. If the explicit score $\nabla(y_t|\theta_{t|t-1})$ is non-zero, a learning rate large enough—or, equivalently, a penalty

small enough—generally exists that leads to the undesirable outcome $p(y_t|\theta^{\text{ex}}_{t|t}) < p(y_t|\theta_{t|t-1})$ due to "overshooting". In contrast, any non-infinitesimal implicit update strictly improves the local fit of the model, for any learning rate, by design. This is because the penalty at $\theta_{t|t}$ strictly exceeds the zero penalty at $\theta_{t|t-1}$. Therefore, the likelihood difference must at least match this non-zero increase in penalty; otherwise $\theta_{t|t}$ is not the optimizer of (1).

The update direction suggested by the explicit score need not be the direction that, for a given step size, achieves the highest likelihood improvement; in fact, it may be detrimental. Generally speaking, the implicit and explicit gradients may point in opposite directions. A special case is when the maximization problem (2) is concave, in which case both strategies suggest adjustments of the time-varying parameter that point roughly in the same direction. Geometrically, the angle between the difference vector $\theta_{t|t} - \theta_{t|t-1}$ and the explicit score $\nabla(y_t|\theta_{t|t-1})$ cannot exceed 90 degrees.

Proposition 1 (Relationship between ISD and ESD updates) *Fix $t > 0$ and let Assumptions 1, 2 and 4a hold. Consider a prediction $\theta_{t|t-1} \in \Theta$ and positive-definite penalty $P_t \in \mathbb{R}^{K\times K}$. Compute $\theta_{t|t}$ using the update step (1). Then, with probability one,*

$$\langle \theta_{t|t} - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1}) \rangle \;\geq\; 0. \tag{8}$$

If Assumptions 3 and 4b also hold, we may write:

$$\theta_{t|t} = \theta_{t|t-1} + (P_t + \mathcal{I}_{t|t})^{-1} \nabla(y_t|\theta_{t|t-1}), \tag{9}$$

where $\mathcal{I}_{t|t}$ denotes the negative average $K \times K$ Hessian between $\theta_{t|t-1}$ and $\theta_{t|t}$,

$$\mathcal{I}_{t|t} := -\int_0^1 \left. \frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'} \right|_{\theta = u\,\theta_{t|t-1} + (1-u)\,\theta_{t|t}} \mathrm{d}u. \tag{10}$$

For a scalar time-varying parameter (i.e., $K = 1$), equation (8) implies that the implicit and explicit parameter adjustments—$\theta_{t|t} - \theta_{t|t-1}$ and $\theta^{\text{ex}}_{t|t} - \theta_{t|t-1}$—have the same sign. Naturally, the implicit and explicit gradients—$\nabla(y_t|\theta_{t|t})$ and $\nabla(y_t|\theta_{t|t-1})$—likewise have the same sign. For the ISD update, it follows that the derivative of $\log p(y_t|\theta)$ evaluated at the update has the same sign as the derivative at the prediction. Because the derivative cannot switch signs, the ISD update increases the value of $\log p(y_t|\theta)$ without overshooting. For the ESD update, in contrast, the derivative of $\log p(y_t|\theta)$ at the update may have the opposite sign from the derivative at the prediction. As a result, the explicit update may surpass the peak of $\log p(y_t|\theta)$. This possibility of overshooting means the log-likelihood value at the ESD update may be inferior to that at the prediction.

In a multi-parameter setting (i.e., $K > 1$), the direction of the updates of some elements of the parameter vector can differ between the implicit and explicit updates. This is because the ESD update does not account for interaction effects between the updates of different elements of the parameter vector on the local likelihood. That is, all interactions stem entirely from the learning rate matrix H_t; if it is diagonal, the joint ESD update is akin to updating each individual parameter separately. By preserving the full likelihood function, the ISD update accounts for the interplay between parameters.

The second result of Proposition 1 shows that the ISD update can be written as a "curvature-corrected" version of the ESD update. Specifically, $\mathcal{I}_{t|t}$ is the average negative $K \times K$ Hessian between $\theta_{t|t-1}$ and $\theta_{t|t}$, measuring the average curvature of the postulated logarithmic observation density $\log p(y_t|\theta)$ between these points. As a result, the ISD method is better able to control the magnitude of the update as it accounts for the (second-order) impact on the local likelihood. If the log likelihood is linear in θ, then $\mathcal{I}_{t|t} = O_K$, such that the ISD and ESD updates are identical; this is logical as it means the linear approximation in the ESD update is exact. If the log likelihood is (multivariate) quadratic in θ—as with the normal distribution in terms of the mean—then $\mathcal{I}_{t|t}$ is constant. In that case, the ISD and ESD updates are equivalent, albeit for different penalty matrices.

To say more about the properties of the ISD update (1), we need more information about the shape of the log-likelihood function $\log p(y_t|\theta)$. In this paper, we focus on the family of concave log-likelihood functions, which allows us to derive a set of particularly strong global optimality and stability properties.

Assumption 5 (Log-concave observation density) $\log p(y_t|\theta) + \alpha_t/2 \, \|\theta\|^2$ *is concave in* θ *for some* $\alpha_t \geq 0$, $\forall \theta \in \Theta$, *with probability one.*

Assumption 5 is a stronger version of Assumption 2, as it imposes concavity on the log-likelihood contribution itself, rather than on its regularized version (2). The strength of concavity is measured by $\alpha_t \geq 0$, where the boundary case $\alpha_t = 0$ implies concavity while $\alpha_t > 0$ implies α_t-strong concavity. Many popular logarithmic densities are concave in their parameters, as illustrated in our simulations and empirical analysis. While Assumption 5 yields strong theoretical results, the optimization literature suggests that implicit gradient methods remain effective in practice if the logarithmic density fails to be concave (e.g., Hare and Sagastizábal, 2009). In such settings, the global nature of the ISD update (1) is likely to further enhance its advantages relative to explicit methods (e.g., Grimmer et al., 2023), as confirmed in our simulation studies (Section 5).

Under Assumption 5, and when both methods use the same penalty matrix P_t, the implicit gradient update is a "shrunken" version of the explicit gradient update.

Proposition 2 (Step-size shrinkage) *Fix $t > 0$ and let Assumptions 1 to 5 hold. Take a prediction $\theta_{t|t-1} \in \Theta$ and positive-definite penalty $P_t \in \mathbb{R}^{K\times K}$ as given. Based on the observation y_t, compute $\theta_{t|t}$ using the implicit update (1) and $\theta^{\text{ex}}_{t|t}$ using the explicit update (5). Let $\lambda_{\max}(P_t)$ denote the largest eigenvalue of P_t. Then, with probability one,*

$$\left\|\theta_{t|t} - \theta_{t|t-1}\right\|^2_{P_t} \leq \underbrace{\left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2}_{\in [0,1],\ \textit{contraction coefficient}} \left\|\theta^{\text{ex}}_{t|t} - \theta_{t|t-1}\right\|^2_{P_t}. \tag{11}$$

The contraction coefficient depends on the ratio between the strength of concavity α_t and the penalty P_t, where a larger α_t or smaller $\lambda_{\max}(P_t)$ implies more shrinkage. This same contraction coefficient reappears in both the stability and the optimality results of Section 3. Intuitively, for a concave log-likelihood, every tangent line lies above the curve, such that the explicit method (based on a linear approximation) *over*estimates the likelihood gain that can be achieved by updating. As such, the larger magnitude of the explicit step size can be attributed to "overshooting".

In practice, the shrinkage of the vector $\theta_{t|t} - \theta_{t|t-1}$ evident from equation (11) provides an additional level of robustness that is particularly useful for dealing with outliers. This shrinkage property enables ISD filters to use larger learning rates, making them considerably more (rather than less) responsive.

3 Theory

3.1 Stability

Turning to the stability properties of the proposed framework, we are particularly interested in providing sufficient conditions for filter invertibility, meaning that filtered paths based on identical data but with different initializations converge exponentially fast over time. We remain agnostic with regard to the DGP and use assumptions relating to the filter only.

Our results in this section are presented in three parts: (a) fixing t and examining only the update step (Lemma 1), (b) fixing t and considering both the update and prediction steps (Lemma 2), and (c) proving invertibility by considering the composition of all prediction-to-prediction mappings (Theorem 1).

We begin by fixing the time step t and evaluating the stability of the update step. Lemma 1 shows that the ISD update (1) is stable under Assumptions 1 through 5, while, absent further conditions, the same does not hold for the ESD update (6).

Lemma 1 (Prediction-to-update stability) *Fix $t > 0$ and let Assumptions 1 to 5 hold. Let $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$ denote two predictions in Θ, which are combined with the observation y_t in the ISD update step (1) to yield corresponding parameter updates, $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$. Then, with probability one,*

$$\left\|\theta_{t|t} - \tilde{\theta}_{t|t}\right\|_{P_t}^2 \leq \underbrace{\left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2}_{\in [0,1],\ \textit{contraction coefficient}} \left\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\right\|_{P_t}^2, \tag{12}$$

where $\lambda_{\max}(P_t)$ is the largest eigenvalue of P_t. For the ESD update (6), under the additional assumptions that $\nabla(y_t|\theta)$ is L_t-Lipschitz continuous in θ with probability one and $\lambda_{\min}(P_t) \geq L_t/2$, where $\lambda_{\min}(P_t)$ is the smallest eigenvalue of P_t, with probability one,

$$\left\|\theta_{t|t}^{\text{ex}} - \tilde{\theta}_{t|t}^{\text{ex}}\right\|_{P_t}^2 \leq \underbrace{\frac{\lambda_{\max}(P_t) - \alpha_t[2 - L_t/\lambda_{\min}(P_t)]}{\lambda_{\max}(P_t)}}_{\in [0,1],\ \textit{contraction coefficient}} \left\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\right\|_{P_t}^2. \tag{13}$$

The first result of Lemma 1 demonstrates that the ISD update step is non-expansive in the squared norm $\|\cdot\|_{P_t}^2$; that is, it does not magnify (and possibly shrinks) the distance between different paths. For a strongly concave log-likelihood function (i.e., $\alpha_t > 0$), we obtain a strict contraction in the norm $\|\cdot\|_{P_t}$ as long as the predictions are not identical (i.e., $\theta_{t|t-1} \neq \tilde{\theta}_{t|t-1}$). The strength of the contraction is determined by the strength of concavity α_t and the maximum eigenvalue of the penalty matrix P_t.

The second result of Lemma 1 shows that a similar non-expansiveness result can be obtained for the ESD update (6), but this requires two additional assumptions. Namely, the score needs to be L_t-Lipschitz continuous in θ (or, equivalently, $\log p(y_t|\theta)$ needs to be L_t-smooth) and the penalty matrix P_t must exceed $L_t/2$ in an eigenvalue sense. The latter condition is equivalent to saying that H_t must be exceeded by $2/L_t$ in an eigenvalue sense. That the learning rate H_t must shrink as the Lipschitz constant L_t increases is well known in the SGD literature (e.g., Karimi et al., 2016), but is, to our knowledge, new in the ESD literature. This insight is crucial for understanding the potential instability of ESD filters in the absence of L-smoothness, as illustrated in our simulation studies (Section 5).

We now turn to the prediction-to-prediction mapping from time step t to $t+1$. To obtain a strictly contracting prediction-to-prediction mapping for the ISD filter, it is sufficient for the update and prediction steps be non-expansive in the norm $\|\cdot\|_{P_t}$, providing at least one of them is strictly contractive. That is, when $\alpha_t = 0$, the prediction mapping from $\theta_{t|t}$ to $\theta_{t+1|t}$ must be strictly contracting in the norm $\|\cdot\|_{P_t}$. When $\alpha_t > 0$, on the other hand, it is sufficient for the prediction step to be non-expansive. For example, the identity mapping

$\theta_{t+1|t} = \theta_{t|t}$ is non-expansive and often useful in practice.

A sufficient condition for non-expansiveness (contractiveness) of the prediction step in the norm $\|\cdot\|_{P_t}$ is that $P_t \succeq \Phi' P_t \Phi$ ($P_t \succ \Phi' P_t \Phi$). Here, the notation $X \succeq Y$ ($X \succ Y$) indicates that $X - Y$ has non-negative (strictly positive) eigenvalues for two symmetric real-valued matrices X and Y of the same size. This requirement is equivalent to $\|\Phi\|_{P_t} \leq 1$ ($\|\Phi\|_{P_t} < 1$), where $\|X\|_{P_t}$ is the induced operator norm of a matrix $X \in \mathbb{R}^{K \times K}$, which is closely related to the discrete Lyapunov equation (e.g., Anderson and Moore, 2012). Lemma 2 summarizes the contraction of the prediction-to-prediction mapping in $\|\cdot\|_{P_t}$.

Lemma 2 (Prediction-to-prediction stability) *Fix $t > 0$ and let Assumptions 1 to 5 hold. Let P_t be given with $P_t \succeq \Phi' P_t \Phi$. Let $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$ denote two predictions in Θ that are used in the ISD update step (1) to yield the corresponding parameter updates $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$, and are subsequently passed to the prediction step (4) to yield predictions $\theta_{t+1|t}$ and $\tilde{\theta}_{t+1|t}$. With probability one,*

$$\left\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\right\|_{P_t}^2 \leq \kappa_t \left\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\right\|_{P_t}^2, \tag{14}$$

where the contraction coefficient κ_t is

$$\kappa_t = \frac{\lambda_{\max}(P_t)[\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)]}{(\lambda_{\max}(P_t) + \alpha_t)^2}. \tag{15}$$

If either $\alpha_t > 0$ or $P_t \succ \Phi' P_t \Phi$, then, with probability one, $\kappa_t \in [0, 1)$.

The strength of the contraction of the prediction-to-prediction mapping at time t is measured by κ_t, which is a function of the strength of concavity α_t, the penalty matrix P_t, and the autoregressive matrix Φ. For a scalar time-varying parameter, the standard condition $|\Phi| < 1$ is sufficient to yield $\kappa_t \in [0, 1)$. In the multiple-parameter setting, $\Phi'\Phi \prec I_K$ implies $\Phi' P_t \Phi \prec P_t$ when (a) Φ and P_t are both diagonal or (b) either Φ or P_t is a constant multiple of the identity. In this case, the standard condition $\rho(\Phi) < 1$, where $\rho(\cdot)$ denotes the spectral radius, is sufficient to yield $\kappa_t \in [0, 1)$. To allow for more richly parameterized Φ and P_t, we could express P_t as the solution to the discrete version of Lyapunov's equation $P_t - \Phi' P_t \Phi = \Delta_t \succ 0$, which for a given $\rho(\Phi) < 1$ has a unique solution $P_t \succ 0$ parameterized in terms of $\Delta_t \succ 0$ (see e.g, Bof et al., 2018, Thrm 3.2). The strict inequalities in this paragraph could be weak inequalities if we additionally impose $\alpha_t > 0$.

Finally, we analyze the composition of all prediction-to-prediction mappings. For the effects of the initialization of the filter at time $t = 0$ to disappear exponentially fast, the composition of all prediction-to-prediction mappings is required to be contractive. A sufficient (but stronger than necessary) condition is that each individual prediction-to-prediction

mapping is contractive in a single (i.e., shared) norm across all mappings over time. Theorem 1 formulates sufficient conditions for the existence of such a shared norm and contains an invertibility result that is crucial in enabling maximum-likelihood estimation of the static parameters (e.g., Straumann and Mikosch, 2006). This desirable invertibility property further ensures that numerical errors do not accumulate during implementation in practice—a concern also expressed for the Kalman filter (Anderson and Moore, 2012).

Theorem 1 (Invertibility) *For all $t > 0$, let Assumptions 1 to 5 hold, with either (a) $P_t \succ \Phi' P_t \Phi$ or (b) $P_t \succeq \Phi' P_t \Phi$ and $\alpha_t > 0$. In addition, let there be some $\bar{P}, A \in \mathbb{R}^{K \times K}$ with $\bar{P} \succ A \succ O_{K \times K}$ and a sequence $\{\rho_t > 0\}$ such that for all $t > 0$, with probability one,*

$$\kappa_t P_t + \rho_t A \preceq \rho_t \bar{P} \preceq P_t, \tag{16}$$

where κ_t is defined in (15). Take two initial values $\theta_{0|0} \in \Theta$ and $\tilde{\theta}_{0|0} \in \Theta$, yielding two sequences $\{\theta_{t|t-1}\}$ and $\{\tilde{\theta}_{t|t-1}\}$, respectively. Then the filter composed of (1) and (4) is invertible, *i.e., there exists a constant $c > 1$ such that with probability one,*

$$\lim_{t \to \infty} c^t \left\| \theta_{t|t-1} - \tilde{\theta}_{t|t-1} \right\|^2 \to 0. \tag{17}$$

Theorem 1 expresses a sufficient condition for a contraction of all prediction-to-prediction mappings in the common norm $\|\cdot\|_{\bar{P}}$, where $\bar{P}$ is a time-invariant matrix satisfying inequality (16). For a scalar time-varying parameter, this condition is guaranteed irrespective of the sequence $\{P_t\}$ whenever the standard condition $|\Phi| < 1$ holds. For the unit-root case $|\Phi| = 1$, it is sufficient that $\{P_t\}$ is upper bounded while $\{\alpha_t\}$ is strictly lower bounded away from zero, in both cases uniformly over time, thereby preventing κ_t from approaching unity. In the multiple-parameter setting, equation (16) essentially limits only the relative dynamics of $\{P_t\}$, preventing the penalization of different elements of the time-varying parameter vector from varying too drastically over time. Condition (16) is less stringent when the persistence in the prediction step is reduced (i.e., for Φ closer to $O_{K \times K}$) and/or when the strength of concavity is increased (i.e., for larger $\{\alpha_t\}$), as these conditions lead to stronger contractions (i.e., lower $\{\kappa_t\}$).

The presence of the scalar $\rho_t > 0$ in condition (16) indicates that the relative penalization between parameters matters, but not the overall magnitude. This is because a contraction in the norm $\|\cdot\|_P$ implies a contraction in the norm $\|\cdot\|_{\rho_t P}$ and vice versa. For this reason, condition (16) is automatically satisfied if the sequence $\{P_t\}$ is a time-varying scalar multiple of a static matrix: $\{P_t = \zeta_t P\}$ for some sequence $\{\zeta_t > 0\}$ and $P \succ O_{K \times K}$ for which $P \succ \Phi' P \Phi$. Matrix A in condition (16) is included to ensure that the contraction

coefficient with respect to the norm $\|\cdot\|_{\bar{P}}$ is bounded above, uniformly across time, at some value strictly below unity.

Importantly, Theorem 1 relies on the researcher-postulated density $p(\cdot|\theta_t)$, but not the true observation density $p_0(\cdot|\theta_t^0)$. Invertibility in the ISD framework can thus be guaranteed without imposing additional restrictions on the DGP, which is convenient as this is typically unknown. We may even allow Assumptions 1 to 5 to fail for a particular realization of the observation, as long as this violation occurs with probability zero. When the assumptions are guaranteed to hold for all possible y_t, the above stability result is entirely unaffected by model misspecification. Hence, result (17) implies the exponential almost sure (e.a.s.) convergence of the paths $\{\theta_{t|t-1}\}$ and $\{\tilde{\theta}_{t|t-1}\}$ based on the same data, such that differences stemming from either (a) the varying initializations $\theta_{0|0}$ and $\tilde{\theta}_{0|0}$ or (b) numerical errors due to finite computer precision disappear exponentially fast as time progresses.

In contrast, Lemma 1 indicated that the stability of ESD filters is contingent on the magnitude of the learning rate H_t. By additionally assuming that $\nabla(y_t|\theta)$ is L_t-Lipschitz continuous and $\lambda_{\max}(H_t) \leq 2/L_t$, $\forall t$, we can use Lemma 1 to construct an invertibility result for ESD filters analogous to that in Theorem 1 for ISD filters. To the best of our knowledge, this would be the first general multivariate invertibility result for ESD filters that does not require knowledge of the DGP. For example, the contraction condition in Blasques et al. (2022) uses an expectation with respect to the DGP. The additional constraints they impose on $\{H_t\}$ suggest that ESD filters require careful tuning to ensure stability. Furthermore, as our simulation studies show (Section 5), if L_t-smoothness is violated, no positive learning rate H_t may exist that guarantees non-expansiveness of the ESD update over the entire parameter space Θ. In this case, we must ensure that unstable regions of the parameter space are visited with sufficiently low probability. As a result, the maximum permitted learning rate will typically be tied to the DGP, and infringing on this (unknown) upper bound may cause filter divergence (e.g., Blasques et al., 2018, p. 1023).

In the optimization literature, Toulis and Airoldi (2017) similarly find that explicit gradient methods require finetuning to avoid divergence; this holds even when the target is static. Arguably, the superior stability of implicit methods is even more relevant in our (dynamic) setting because, unlike in optimization, our filter must remain perpetually responsive without ever converging; hence, stability over time is vital. If there is a positive probability that the ESD filter diverges, it eventually will. Theorem 1 guarantees that ISD filters with log-concave postulated densities are stable under easily verifiable conditions that are agnostic about the DGP.

3.2 Optimality

In addition to ensuring filter stability, we are interested in whether the updating mechanism improves the quality of parameter estimation. To this end, we must reintroduce some consideration of the true process. Under misspecification, we can only hope to recover, as accurately as possible, a *pseudo*-true parameter denoted as $\theta_t^\star$. Here we demonstrate that, for any fixed time step, the ISD update globally contracts toward some small region around this pseudo-true parameter. We require the following additional assumptions.

Assumption 6 (Uniqueness of pseudo-truth) *There exists a* $\theta_t^\star$ *such that* $\forall \theta \in \Theta \setminus \{\theta_t^\star\}$, *we have* $\mathop{\mathbb{E}}_{y_t}[\log p(y_t|\theta_t^\star)] > \mathop{\mathbb{E}}_{y_t}[\log p(y_t|\theta)]$ *and* $\mathop{\mathbb{E}}_{y_t}[\nabla(y_t|\theta_t^\star)] = 0$.

Assumption 7 (Bounded information) $\mathop{\mathbb{E}}_{y_t}[\|\nabla(y_t|\theta_t^\star)\|^2] < \infty$.

Assumption 6 asserts the existence of a unique pseudo-truth $\theta_t^\star$ that maximizes the expected (postulated) log-likelihood function $\mathop{\mathbb{E}}_{y}[\log p(y|\theta_t^\star)]$. Equivalently, $\theta_t^\star$ is the unique minimizer of the Kullback-Leibler divergence of $p(\cdot|\theta_t)$ to the true density $p_0(\cdot|\theta_t^0)$. If the logarithmic postulated density is differentiable and strongly concave with probability one—i.e., Assumptions 4a and 5 with $\alpha_t > 0$ hold—then the existence of a unique pseudo-truth is automatic and need not be separately assumed. In the case of correct model specification, the truth and pseudo-truth coincide (i.e., $\theta_t^0 = \theta_t^\star$). Assumption 7 posits that the norm of the squared score computed with the postulated density, and evaluated in the pseudo-truth, is finite in expectation with respect to the true observation density.

If the prediction $\theta_{t|t-1}$ deviates substantially from the pseudo-truth $\theta_t^\star$, leveraging the information from the observation y_t generally yields $\theta_{t|t}$ an improvement upon $\theta_{t|t-1}$. Thus, when the initial prediction is relatively inaccurate, obtaining a more precise parameter estimate is straightforward. However, as $\theta_{t|t-1}$ approaches the pseudo-truth $\theta_t^\star$, further improvement becomes increasingly difficult. For highly accurate predictions, the update $\theta_{t|t}$ may even be less accurate than the prediction due to its reliance on the noisy observation y_t. Indeed, no improvement is possible if the prediction is already exact (i.e., in the case $\theta_{t|t-1} = \theta_t^\star$). This is not a limitation of our approach but an inherent characteristic of stochastic optimization methods. Consequently, the region around the pseudo-truth $\theta_t^\star$ is referred to as the *noise-dominated region* (NDR; e.g., Ryu and Boyd, 2016, p. 15, Patrascu and Necoara, 2018, p. 3, Lange, 2024a, Fig. 1). Intuitively, when a prediction is perfect, the data provide no (additional) information.

Since improvements are not always guaranteed, Theorem 2 explicitly characterizes the tug of war between contractive and expansive forces, which respectively decrease and increase the mean squared error (MSE). Which of the two forces dominates largely depends on the

accuracy of the prediction. While most authors establish upper bounds on the MSE after updating, our equations (18)–(19) below provide exact equalities (rather than inequalities). This approach enables us to identify the precise conditions under which updates lead to improvement (see further discussion below).

Theorem 2 (Contraction to the NDR) *Fix $t > 0$ and let Assumptions 1 to 7 hold. Then, for the ISD update (1), we have*

$$\underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t}-\theta_t^\star\right\|_{P_t}^2\right]}_{\text{MSE after update}} = \underbrace{\left\|\theta_{t|t-1}-\theta_t^\star\right\|_{P_t}^2}_{\text{SE before update}} - \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t}-\theta_t^\star\right\|_{2\mathcal{I}_{t|t}^\star+\mathcal{I}_{t|t}^\star P_t^{-1}\mathcal{I}_{t|t}^\star}^2\right]}_{\geq 0,\ \text{contractive force}} + \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|_{P_t^{-1}}^2\right]}_{\geq 0,\ \text{expansive force}}. \quad (18)$$

For the ESD update (6), we have

$$\underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t}^{\text{ex}}-\theta_t^\star\right\|_{P_t}^2\right]}_{\text{MSE after update}} = \underbrace{\left\|\theta_{t|t-1}-\theta_t^\star\right\|_{P_t}^2}_{\text{SE before update}} - \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t-1}-\theta_t^\star\right\|_{2\mathcal{I}_{t|t-1}^\star}^2\right]}_{\geq 0,\ \text{contractive force}} + \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_{t|t-1})\right\|_{P_t^{-1}}^2\right]}_{\geq 0,\ \text{expansive force}}. \quad (19)$$

Here $\mathcal{I}_{t|t}^\star$, $\mathcal{I}_{t|t-1}^\star \succeq \alpha_t I_K \succeq O_K$ denote the negative average $K \times K$ Hessians between $\theta_{t|t}$ or $\theta_{t|t-1}$ and $\theta_t^\star$, that is,

$$\mathcal{I}_{t|t}^\star := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t}+(1-u)\,\theta_t^\star} \mathrm{d}u, \quad (20)$$

$$\mathcal{I}_{t|t-1}^\star := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t-1}+(1-u)\,\theta_t^\star} \mathrm{d}u. \quad (21)$$

To the best of our knowledge, Theorem 2 is new in the stochastic optimization literature: we have not found equations (18)–(19) in prominent contributions such as Parikh and Boyd (2014), Polson et al. (2015), Ryu and Boyd (2016), Bianchi (2016), and Asi and Duchi (2019). Because Theorem 2 focuses on a single time step, the comparison with the stochastic optimization literature is relevant (recall the discussion in Section 1.1). If the second term on the right-hand side of (18) is removed and the equality replaced with an inequality, we obtain a result similar to that in Theorem 3.2 of Asi and Duchi (2019). Because the expansive force in (18) is bounded by Assumption 7, Asi and Duchi (2019, p. 2264) conclude that implicit updates are "nondivergent", while explicit updates lack this guarantee since the expansive force in (19) is not uniformly bounded. Our contribution in Theorem 2 is the inclusion of contractive forces, which play a crucial role in improving updates over predictions. These terms enable us to express equalities rather than inequalities. Given that updates should ideally be more accurate than predictions (at least outside the NDR), these contractive forces

are key. While Assumption 5 (concavity) ensures their positivity, it stronger than necessary; a weaker condition, $\mathcal{I}^\star_{t|t}, \mathcal{I}^\star_{t|t-1} \succeq O_K$, would suffice. In any case, equations (18) and (19) remain valid, and since our analysis is based on equalities, it cannot be further improved.

Next, we discuss equations (18)–(19) in more detail. On the left-hand side, we have the mean squared error (MSE) of ISD and ESD updates, measured in the $\|\cdot\|_{P_t}$ norm as denoted by $\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t} - \theta^\star_t\right\|^2_{P_t}\right]$ and $\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta^{\text{ex}}_{t|t} - \theta^\star_t\right\|^2_{P_t}\right]$, respectively. The right-hand sides of both equations contain three components. First, the weighted squared error (SE) before the update, $\|\theta_{t|t-1} - \theta^\star_t\|^2_{P_t}$, reflects the deviation of the prediction $\theta_{t|t-1}$ from the pseudo-truth $\theta^\star_t$. The remaining two terms will therefore determine whether, in expectation, the SE is reduced after updating.

Second, we have the contractive factor, which differs between the ISD and ESD updates. For the ISD update, the contractive force in (18) is the MSE after updating using the norm $\|\cdot\|_{2\mathcal{I}^\star_{t|t} + \mathcal{I}^\star_{t|t} P_t^{-1} \mathcal{I}^\star_{t|t}}$, while for the ESD update in (19) it is the SE before updating using the norm $\|\cdot\|_{2\mathcal{I}^\star_{t|t-1}}$. In both cases, the magnitude of the contractive force is proportional to the strength of concavity measured by $\mathcal{I}^\star_{t|t}$ and $\mathcal{I}^\star_{t|t-1}$, which are the negative average Hessians between the updated parameter $\theta_{t|t}$ or the predicted parameter $\theta_{t|t-1}$ and the pseudo-truth $\theta^\star_t$. This suggests that, to obtain a contractive force, we do not necessarily need the postulated log likelihood $\log p(y_t|\theta)$ to be concave for all $\theta \in \Theta$. Hence Assumption 5 is essentially too strong. All that is required is, on average, sufficient curvature in the direction of the pseudo-truth $\theta^\star_t$. Effectively, these curvature conditions ensure that the gradient is on average pointing in the correct direction, while its magnitude increases sufficiently rapidly as we move away from the pseudo-truth. The contractive force is contained within the expectation operator $\mathop{\mathbb{E}}_{y_t}[\cdot]$, which uses the true density. In principle, concavity could be allowed to fail for certain realizations y_t, as long as these do not occur with high probability. Similar versions of strong concavity can be found in the optimization literature (e.g., Toulis et al., 2021, Assumption 3).

Third, we have the expansive force, which reveals important differences between the ISD and ESD updates. For the ISD update, the expansive force is $\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta^\star_t)\right\|^2_{P_t^{-1}}\right]$, i.e., the weighted norm of the postulated gradient evaluated at the pseudo-truth and averaged over y_t using the true density. This term reflects the irreducible noise obtained by updating based on the noisy observation y_t. Importantly, the magnitude of the irreducible noise is *in*dependent of the prediction $\theta_{t|t-1}$; hence, the strength of the expansive force remains constant as $\theta_{t|t-1}$ moves further from the pseudo-truth $\theta^\star_t$. In contrast, the strength of the contractive force of the ISD update increases with the distance of $\theta_{t|t}$ from $\theta^\star_t$. As a result, it will dominate the irreducible noise when $\theta_{t|t}$ is far from $\theta^\star_t$. In the region where this contractive force dominates (i.e., outside the NDR), we expect updates to be beneficial.

For the ESD update in (19), the expansive force is the expected weighted norm of the explicit gradient, i.e., $\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_{t|t-1})\right\|^2_{P_t^{-1}}\right]$, which crucially depends on $\theta_{t|t-1}$. For log-concave distributions, this term will also increase with the distance of $\theta_{t|t-1}$ from θ_t^*. Without further assumptions on the postulated density and/or DGP, it is unclear whether, for bad predictions, the expansive or contractive force dominates. If the expansive force dominates for predictions far from the pseudo-truth, the filter may diverge. In our simulation studies, we will see that this happens frequently. Furthermore, both the contractive and expansive forces of the ISD update are increasing in $H_t = P_t^{-1}$ (provided $\mathcal{I}^\star_{t|t} \succ O_K$), while for the ESD update this holds only for the expansive force. This highlights the importance of selecting the correct learning rate for the ESD method. In particular, unless $\nabla(y_t|\theta_{t|t-1}) = 0$ with probability one, we can make the MSE after the ESD update arbitrarily bad by the letting the learning rate tend to infinity. This is because the ESD method can be prone to overshooting. While we may use Lipschitz-gradient continuity combined with a stringent learning-rate restriction to demonstrate a contraction to the NDR (see online Supplement B.1 for details), the main message here is that, without further restrictions, ESD updates are *not* always beneficial; this observation goes against a large body of literature as indicated in Table 1.

Because the expectation operator $\mathop{\mathbb{E}}_{y_t}[\cdot]$ uses the unknown true density, it is generally difficult to pinpoint the magnitude of the contractive and expansive forces. In particular, the analysis is complicated by the dependence of the curvature strength $\mathcal{I}^\star_{t|t}$ on y_t. If the log likelihood $\log p(y_t|\theta)$ is strongly concave with probability one (i.e., Assumption 5 with $\alpha_t > 0$ holds), we may obtain a particularly strong—indeed, much stronger than necessary—type of contraction of the ISD update toward the NDR.

Corollary 1 (Geometric contraction to the NDR) *Fix $t > 0$ and let Assumptions 1 to 7 hold, where Assumption 5 holds for some $\alpha_t > 0$. Then the ISD update (1) satisfies*

$$\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t} - \theta_t^\star\right\|^2_{P_t}\right] \leq \underbrace{\left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2}_{\in [0,1),\ \textit{contraction coefficient}} \left(\left\|\theta_{t|t-1} - \theta_t^\star\right\|^2_{P_t} + \mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|^2_{P_t^{-1}}\right]\right). \tag{22}$$

Equation (22) displays a global linear contraction toward the pseudo-truth $\theta_t^\star$ up to some level of accuracy determined by the expected squared norm of the score at $\theta_t^\star$. This contraction is geometric in the sense that, for large prediction errors, the ratio of the MSE after updating to the SE before updating is equal to a contraction coefficient less than unity. This contraction coefficient is the same as in Proposition 2 and Lemma 1 and is regulated by the curvature of the log-likelihood function, measured by α_t, relative to the size of the penalty measured by $\lambda_{\max}(P_t)$. *Ceteris paribus*, a smaller penalty or stronger concavity

therefore yields a faster contraction. Note that by leveraging the equality in Theorem 2, we obtain a contraction that is stronger than comparable results in the literature (e.g., Lange, 2024a, Thrm 1). As the prediction $\theta_{t|t-1}$ approaches the pseudo-truth $\theta_t^\star$, in which case further improvement is impossible, we see that the MSE after updating is controlled by the penalty matrix P_t and the irreducible noisiness of the observations, which are combined in the quantity $\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|^2_{P_t^{-1}}\right]$. This quantity decreases as the penalty P_t increases, such that larger learning rates (or, equivalently, smaller penalties) lead to a larger NDR. The optimal choice of learning rate is therefore determined by a trade-off between contraction speed when far from the pseudo-truth and the size of the NDR. As we discuss in the next section, the optimal penalty can be estimated with maximum likelihood using available data.

Finally, similar to Lange (2024a, Prop. 2), we can apply the geometric contraction in Corollary 1 to provide an upper bound on the long-run MSE of $\theta_{t|t}$ relative to $\theta_t^\star$. However, doing so will require placing additional restrictions on the dynamics of the true (or pseudo-true) parameter, which we leave for future work.

4 Estimation

The parameters of the ISD filter, including the penalty matrices $\{P_t\}$ in the update (1), parameters ω and Φ in the prediction step (4), and any additional static shape parameters $\psi \in \Psi \subseteq \mathbb{R}^M$ in the observation density, are generally unknown and need to be estimated. In our simulations and empirical illustrations, the penalty matrix is taken to be constant: $P_t = P$ for all t. (While we could set P_t proportional to a power of the information matrix as in the ESD literature (e.g., Creal et al., 2013), we do not pursue this here.) The static parameters can be determined through maximum-likelihood (ML) estimation based on the standard prediction-error decomposition, which is also the de facto standard for ESD models (e.g., Creal et al., 2013). That is, we consider

$$\hat{\xi} := \operatorname*{argmax}_{\xi \in \Xi} \sum_{t=1}^{T} \log p(y_t|\theta_{t|t-1}, \psi), \tag{23}$$

where $\xi := [\text{vech}(P)', \omega', \text{vec}(\Phi)', \psi']'$ is a column vector that stacks all static model parameters, and $\text{vec}(\cdot)$ and $\text{vech}(\cdot)$ are the vectorization and half-vectorization matrix operations. The right-hand side of (23) depends on ξ both directly, via ψ, and indirectly, via the predicted parameter path $\{\theta_{t|t-1}\}$, which itself depends on all elements of ξ. The optimization domain Ξ in (23) is the subset of $\mathbb{R}^{3/2K(K+1)+M}$ for which the penalty matrix is positive defi-

nite ($P \succ O_K$) and the shape parameters are in their respective domains ($\psi \in \Psi$).[2] For large K, it may convenient to use the discrete Lyapunov equation (i.e., $P - \Phi' P \Phi = \Delta \succ O_K$) to reparameterize P as indicated in Section 3.1. This leaves the standard parameter restrictions $\Delta \succ O_K$ and $\rho(\Phi) < 1$, yielding stability via Theorem 1.

Using the results obtained in Blasques et al. (2022) for ESD models, we conjecture that $\hat{\xi}$ is a consistent and asymptotic normally distributed estimator of the pseudo-true $\xi^\star$ under standard regularity conditions, whereby $\xi^\star$ minimizes the Kullback-Leibler divergence to the truth. These conditions include the assumptions that ξ is identified, the series $\{y_t\}$ is stationary ergodic and near-epoch dependent with some finite moments, and the postulated density $p(y|\theta, \psi)$ is sufficiently smooth in its arguments and has bounded derivatives. The latter conditions provide sufficient moments to be used in the appropriate law of large numbers and central-limit theorem (see Blasques et al., 2022, Theorems 4.6 and 4.15 for details).

A further crucial ingredient of the proofs in Blasques et al. (2022) is the invertibility concept posited by Bougerol (1993) and Straumann and Mikosch (2006). Verifying the relevant contraction condition can be challenging for ESD models, as the maximum stable learning rate may depend on the unknown true distribution. ISD models are, on the other hand, far more stable due to their non-overshooting properties and robustness against misspecification of the learning rate $H = P^{-1}$, as demonstrated in Section 3. Theorem 1 presents a particularly strong form of invertibility for ISD models with concave logarithmic observation densities, requiring neither Lipschitz-gradient continuity of the postulated logarithmic density nor knowledge of the true distribution. All empirical examples satisfy the conditions of Theorem 1, while the simulations also contain an example showing ML estimation to be effective for a non-concave log density (Section 5.2). For log-concave ISD models, we conjecture that even multidimensional cases ($K > 1$) should pose no problem for ML estimation. A full asymptotic investigation is left for future research.

5 Simulation studies

In this section, we perform a number of simulation experiments to explore the differences between the ISD filter as proposed in this article and the standard ESD version. The various observation densities are chosen to illustrate specific advantages of the ISD filter over its ESD counterpart. Unless noted otherwise, we simulate 1000 series $\{y_t\}$ of length T, where

[2] For the initialization of the time-varying parameter at time $t = 0$ and identification of a sensible starting point for the shape parameters ψ, we compute $[\hat{\theta}'_{0|0}, \hat{\psi}']' := \underset{\theta \in \Theta,\, \psi \in \Psi}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(y_t|\theta, \psi)$. Alternatively, $\theta_{0|0}$ could be added to ξ in (23).

Table 2: MSEs of filtered states with dynamic Poisson distribution.

		$\sigma_\eta = 0.10$	0.15	0.20	0.25	0.30
ISD filter	In-sample	0.08	0.13	0.20	0.27	0.35
	Out-of-sample	0.08	0.14	0.20	0.29	0.36
ESD filter	In-sample	0.08	0.13	0.19	0.25	0.35
	Out-of-sample	0.09	0.14	0.20	∞	∞

the first R observations are used for estimation of the static model parameters. To evaluate the performance of the ISD and ESD models, we compute the MSEs of the filtered values $\{\theta_{t|t}\}$ and $\{\theta^{\text{ex}}_{t|t}\}$ relative to the true values $\{\theta^0_t\}$. We distinguish between in-sample and out-of-sample MSE, based on observations $t = 1, \ldots, R$ and $t = R+1, \ldots, T$, respectively.

5.1 Dynamic Poisson distribution: Non-Lipschitz gradient

For each time t, $y_t \in \mathbb{N}$ is drawn from a Poisson distribution with a time-varying intensity $\lambda^0_t := \exp(\theta^0_t)$, i.e., $p(y_t|\theta^0_t) = (\lambda^0_t)^{y_t} \exp(-\lambda^0_t)/y_t!$. The score with respect to the log-intensity parameter $\theta \in \mathbb{R}$ is $y_t - \exp(\theta)$, which is clearly non-Lipschitz. The negative Hessian and Fisher information are both $\exp(\theta) > 0$; hence, the density is log-concave in θ. We specify the state dynamics as $\theta^0_t = 0.98\theta^0_{t-1} + \eta_t$, where $\eta_t \overset{\text{i.i.d.}}{\sim} \text{N}(0, \sigma^2_\eta)$, and we vary the value of σ_η.

We consider ISD and ESD filters based on the (correctly specified) Poisson distribution, with $T = 2000$ and $R = 1000$. For the ESD filter, we follow Koopman et al. (2016) by using $\theta^{\text{ex}}_{t|t} = \theta^{\text{ex}}_{t|t-1} + H_t(y_t - \exp(\theta^{\text{ex}}_{t|t-1}))$, where the time-dependent learning rate $H_t > 0$ scales with the inverse square root of the predicted Fisher information quantity; hence, $H_t = H \exp(-\frac{1}{2}\theta^{\text{ex}}_{t|t-1})$ with $H > 0$. When interpreting the results below, note that the difference $\theta^{\text{ex}}_{t|t} - \theta^{\text{ex}}_{t|t-1} = H_t(y_t - \exp(\theta^{\text{ex}}_{t|t-1}))$ fails to be Lipschitz in $\theta^{\text{ex}}_{t|t-1}$, growing exponentially to positive or negative infinity as $\theta^{\text{ex}}_{t|t-1} \to -\infty$ or $\theta^{\text{ex}}_{t|t-1} \to \infty$, respectively. Sizable prediction errors may therefore lead to even larger filtering errors, and vice versa, potentially causing divergence of the ESD filter. This issue cannot be addressed by using a different (or static) learning rate. For example, when a static learning rate $H_t = H$ is used, it remains the case that $H(y_t - \exp(\theta^{\text{ex}}_{t|t-1}))$ fails to be Lipschitz continuous in $\theta^{\text{ex}}_{t|t-1}$. Indeed, the lack of Lipschitz continuity is attributable not to the learning-rate specification, but rather to the highly nonlinear (exponential) link function.

Table 2 shows comparable in-sample MSEs; for both filters, the MSE increases in line with the state variability σ_η. In contrast, the out-of-sample filtering performance is similar only for values of σ_η up to $\sim$0.20. For larger values of σ_η, the out-of-sample MSE of the ESD filter diverges to infinity, while the performance of the ISD filter remains in line with the in-sample results. Larger innovations in the true process may give rise to larger prediction errors; for $10 - 20\%$ of replications, the ESD filter diverged. Koopman et al. (2016) used

Table 3: MSEs of filtered states with dynamic GED.

		$\beta = 0.5$	1	1.5	2	2.5	3	3.5	4
ISD filter	In-sample	0.63	0.64	0.59	0.59	0.58	0.59	0.60	0.60
	Out-of-sample	0.70	0.68	0.61	0.60	0.60	0.60	0.61	0.61
ESD filter	In-sample	10.86	10.93	0.68	0.60	0.66	0.76	0.94	1.18
	Out-of-sample	63.18	175.40	0.70	0.61	0.66	0.78	∞	∞

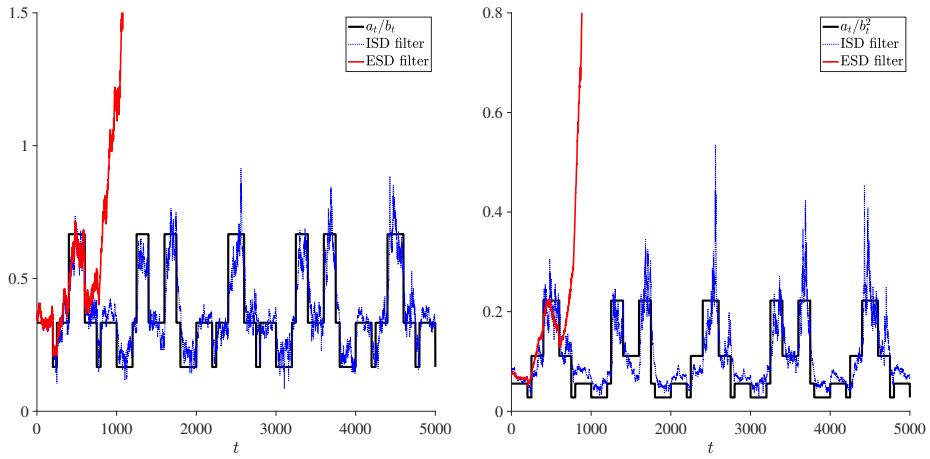
$\sigma_\eta = 0.15$ such that the potential instability of the ESD filter went unnoticed. For the ISD filter, Theorem 1 ensures stability.

5.2 Dynamic GED: Non-concave and non-Lipschitz gradient

For each time t, the observation $y_t \in \mathbb{R}$ is drawn from a generalized error distribution (GED) with a time-varying mean parameter θ_t^0, i.e., $p^0(y_t|\theta_t^0) = \upsilon \exp(-|(y_t - \theta_t^0)/\sigma|^\upsilon)/(2\sigma\Gamma(\upsilon^{-1}))$, where $\Gamma(\cdot)$ denotes the Gamma function, and $\sigma, \upsilon > 0$ are static shape parameters. Here, σ is related to the scale but does not equal the standard deviation. In particular, we take $\sigma^2 = \Gamma(\upsilon^{-1})\Gamma(3\upsilon^{-1})$, which ensures that, conditional on θ_t^0, the variance of y_t is unity. We vary the value of υ, where $\upsilon = 1$ and $\upsilon = 2$ correspond to the Laplace and Gaussian distributions, respectively. For $\upsilon > 1$, the log-density is continuously differentiable and concave in $\theta \in \mathbb{R}$; for $\upsilon < 1$, it is neither. The gradient is Lipschitz only if $\upsilon \in (1, 2]$. A comparative advantage of the ISD filter is that, for every time step, the implicit update $\theta_{t|t}$ must lie between $\theta_{t|t-1}$ and y_t. In contrast, the ESD filter may "overshoot" in that $\theta_{t|t}^{\text{ex}}$ and $\theta_{t|t-1}^{\text{ex}}$ can lie on opposite sides of y_t.

The true parameter evolves as $\theta_t^0 = 0.98\theta_{t-1}^0 + \eta_t$, where $\eta_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$. The signal-to-noise ratio is around one, as the variance of the state innovations equals that of the observation noise. We simulate series with $T = 2000$ and use the first $R = 1000$ observations to estimate the autoregressive parameter $\Phi \in (-1, 1)$ in the prediction step (4) and the learning rate $H > 0$. For simplicity, we use $\omega = 0$ and set υ equal to the true value in the DGP.

Table 3 shows the in- and out-of-sample MSEs of both filters. While the performance for $\upsilon = 2$ is near identical (in which case, the two filters are equivalent), the ISD filter generally achieves substantially lower MSEs. For $\upsilon = 3.5$ and 4, the ESD filter diverges in the out-of-sample period in approximately 25% and 50% of cases, respectively. This is because the gradient is roughly a polynomial of degree $\upsilon - 1$ in the prediction error; that is, it is excessively large for inaccurate predictions. Inaccurate predictions lead to inaccurate updates and vice versa; hence, the filter may diverge. For $\upsilon < 1$, on the other hand, the gradient is unbounded only for highly accurate predictions, which occur less frequently and not consecutively; even as the filter does not diverge, the resulting MSEs are very large indeed.



(a) True and filtered mean of y_t (b) True and filtered variance of y_t

Figure 1: Illustration of filtering performance with dynamic Gamma distribution.

5.3 Dynamic Gamma distribution: Two time-varying parameters and non-smooth DGP

For each time t, an observation $y_t > 0$ is drawn from a Gamma distribution with two dynamic parameters $a_t > 0$ and $b_t > 0$, which are collected in the (true) state vector $\theta_t^0 = (a_t, b_t)' \in \mathbb{R}^2_{>0}$, such that $p(y_t|\theta_t^0) = (b_t)^{a_t} y_t^{a_t - 1} \exp(-b_t y_t)/\Gamma(a_t)$. The same parametrization is used in Fearnhead and Meligkotsidou (2004, eq. 3), albeit in a static context. Conditional on θ_t^0, the mean and variance of y_t are a_t/b_t and a_t/b_t^2, respectively. The ISD filter can be applied directly to $\theta_t = (a_t, b_t)'$, because optimization (1) guarantees that both elements remain positive: the optimization domain Θ is the positive quadrant in $K = 2$ dimensions. For each y_t, the Gamma log-density is (jointly) concave in $(a_t, b_t) \in \mathbb{R}^2_{>0}$; hence, Theorem 1 guarantees ISD filter stability. For the ESD filter, on the other hand, positivity of the time-varying parameters must be enforced through exponential link functions. The resulting log density is neither concave nor L-smooth in the transformed (i.e., logarithmic) parameters, meaning no theoretical guarantees can be made for the ESD filter.

Figure 1 shows filtering results for a single time series $\{y_t\}$ of length $T = 5000$ based on the (non-smooth) DGP $a_t = 2 + \text{sign}\{\sin(2\pi t/400)\}$ and $b_t = 8 + 3\text{sign}\{\cos(2\pi t/1000)\}$. Only the first $R = 500$ observations are used to estimate the static parameters: ω, Φ (assumed diagonal), and P. The figure shows the theoretical mean of the data, a_t/b_t, and its variance, a_t/b_t^2, along with filtered versions. The in-sample performance is roughly similar for both filters, but early in the out-of-sample period, the ESD filter diverges while the ISD filter continues to track the mean and variance of the data relatively well.

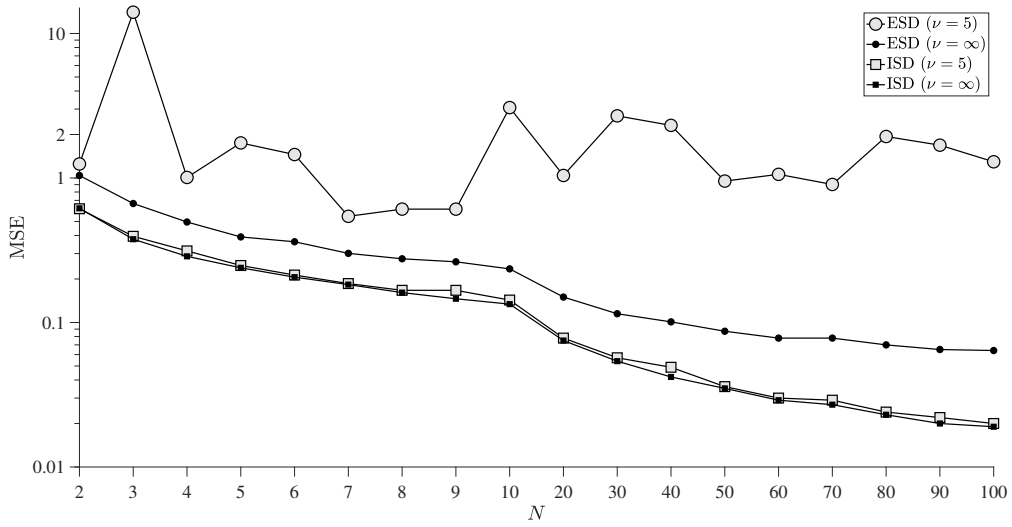


Figure 2: Performance of ISD and ESD filters with N-dimensional observations for dynamic Dirichlet distribution with Gaussian or Student's t state innovations.

5.4 Dynamic Dirichlet: High-dimensional observation and fat-tailed state increments

For each time t, an N-dimensional observation $y_t \in [0,1]^N$ is drawn from a homogeneous Dirichlet distribution with density $p(y_t|\lambda_t^0) = \Gamma(\lambda_t^0 N)/\Gamma(\lambda_t^0)^N \prod_{i=1}^N y_{it}^{\lambda_t^0 - 1}$, where $\lambda_t^0 = \exp(\theta_t^0) > 0$ is the dynamic concentration parameter and $\Gamma(\cdot)$ is the Gamma function. The N elements of y_t are positive and sum to unity (i.e., $y_{it} \geq 0, \forall i$ and $\sum_{i=1}^N y_{it} = 1$). The logarithmic density is concave in θ_t^0, but the gradient is not Lipschitz; hence, no stability guarantees can be made for the ESD filter. The true process satisfies $\theta_t^0 = \omega + \phi\, \theta_{t-1}^0 + \sigma_\eta\, \eta_t$, where the i.i.d. state increments $\{\eta_t\}$ have variance one and are either Gaussian or Student's t distributed with $\nu = 5$ degrees of freedom. The static parameters $\omega = 0.1$, $\phi = 0.95$, and $\sigma_\eta^2 = 0.195$ are chosen such that the unconditional mean and variance of θ_t^0 equal two: $\mathbb{E}(\theta_t^0) = \mathrm{Var}(\theta_t^0) = 2$. We vary the cross-sectional dimension N between 1 and 100. For each N, we simulate 100 series of length $T = 5000$, where the first $R = 500$ observations are used for estimating the static parameters (i.e., ω, ϕ, P).

Figure 2 shows the MSEs of ISD- and ESD-filtered states in the out-of-sample period under both fat-tailed and Gaussian state innovations. In both settings and for any N, the ISD filter outperforms the ESD filter. Notably, the performance of the ISD filter barely deteriorates under fat-tailed (as opposed to Gaussian) state innovations. In contrast, the ESD filter under Student's t innovations is unstable and, unlike the ISD filter, its performance does not consistently improve as the dimensionality N grows. In particular, the MSE of the ESD filter occasionally exceeds the unconditional variance $\mathrm{Var}(\theta_t^0) = 2$. It seems that

abrupt state changes in combination with a non-Lipschitz gradient prohibit the ESD filter from exploiting the increased information content of higher-dimensional draws.

6 Empirical illustrations

6.1 Linear regression with time-varying slope

The capital asset pricing model (CAPM), an important benchmark in finance, links the expected excess returns of individual assets to those of the market in a linear fashion. However, empirical evidence (e.g., Jagannathan and Wang, 1996) shows that the assumption of a constant market coefficient β may be unrealistic, especially in equity markets. Here, we examine the possible time-varying nature of the CAPM market β using the ISD framework. We model the excess asset return y_t as

$$y_t \;=\; \alpha \;+\; \beta_t\, x_t \;+\; \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \tag{24}$$

where α is a static intercept, x_t denotes the excess market return at time t, and ε_t is an i.i.d. normally distributed shock with mean zero and variance σ^2. The ISD update (1) at time t applied to a prediction $\beta_{t|t-1}$ for β_t in (24) can be solved analytically (see online Supplement B.2 for details). Specifically, we obtain the following closed-form solution for $\beta_{t|t}$:

$$\beta_{t|t} \;=\; \beta_{t|t-1} \;+\; \frac{\sigma^2}{\sigma^2 + H x_t^2}\, H\, \nabla(y_t|\beta_{t|t-1}, x_t), \tag{25}$$

where $H = P^{-1} > 0$ is a constant scalar learning-rate parameter and $\nabla(y_t|\beta_{t|t-1}, x_t) := x_t(y_t - \beta_{t|t-1}\, x_t)/\sigma^2$ denotes the explicit score (i.e., evaluated in the prediction $\beta_{t|t-1}$). For the prediction step, we use the linear first-order specification (4).

The ISD update (25) illustrates the shrinkage result of Proposition 2 for the linear regression model. The right-hand side of equation (25) features the shrinkage factor $\sigma^2/(\sigma^2 + Hx_t^2) \in (0, 1]$, which is absent (i.e., equal to unity) in the corresponding ESD update. The amount of shrinkage increases with the magnitude of the explanatory variable x_t^2. This might suggest that the ISD update $\beta_{t|t}$ becomes equal to the prediction $\beta_{t|t-1}$ for large x_t, but this is not the case. In fact, for the ISD update (25) it is straightforward to show that, for a fixed y_t, $\beta_{t|t} \to 0$ if $|x_t| \to \infty$. That the shrinkage factor depends on the exogenous variable x_t appears to be distinctive for the ISD version of the model.

Another difference with the ESD filter is that the ISD update (25) remains bounded as the learning rate H grows. This is evident from the fact that H appears not only in front of the score, but also in the denominator of the shrinkage factor. The practical relevance of this

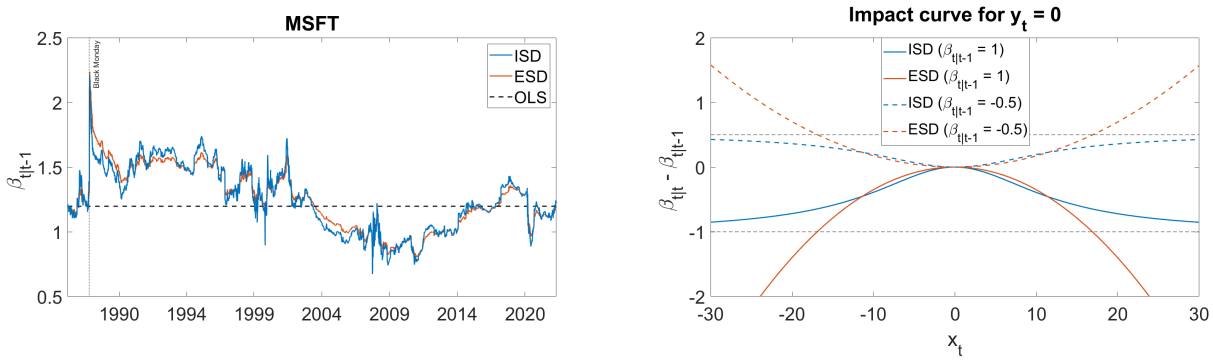


Figure 3: Time-evolution of $\beta_{t|t-1}$ and estimated impact curve of the ISD and ESD models for MSFT from March 1986 until April 2022. Vertical dotted lines mark Black Monday on October 19, 1987.

observation, which extends beyond this linear regression example, is that the ISD filter is robust to the (suboptimal) choice of learning rate, whereas ESD models tend to require more careful finetuning. In the SGD literature, many of the comments above are analogous to those made when comparing the least-mean-square adaptive filter and its normalized version (e.g., Diniz, 1997).

We apply the ISD dynamic regression model (24) to simple daily excess returns of Microsoft (MSFT) from 14 March 1986 to 29 April 2022, obtained from Yahoo Finance.[3] For the market return and risk-free rate, we use the series from Kenneth French's database.[4] Figure 3 shows the evolution of $\beta_{t|t-1}$ for the ISD and ESD models. It also shows the updates $\beta_{t|t} - \beta_{t|t-1}$ as a function of the market return x_t for a fixed $y_t = 0$ and two different predictions ($\beta_{t|t-1} = 1$ and $\beta_{t|t-1} = -0.5$). We refer to these as "impact curves" below.

In Figure 3, the ISD and ESD models produce relatively similar series $\{\beta_{t|t-1}\}$, with a slight advantage for the ISD model. In particular, the ESD model appears to be slow to recover from large shocks, such as the crash on Black Monday, 1987. The reason for this sluggish reaction is that the learning rate H must be substantially reduced to deal with outliers, leading to reduced responsiveness in the remainder of the sample, as is evident around 1994 and 2004.

This problem is drastically reduced for the ISD model by the more favorable (asymptotic) impact curve with respect to the exogenous input. Figure 3 shows an unbounded quadratic impact of x_t on the update $\beta_{t|t} - \beta_{t|t-1}$ in the ESD model, while the ISD impact curve is similar for small $|x_t|$ but bounded for large $|x_t|$. Specifically, for the ISD model, $|x_t| \to \infty$ implies $\beta_{t|t} - \beta_{t|t-1} \to -\beta_{t|t-1}$ and hence $\beta_{t|t} \to 0$. This illustrates the abovementioned property of the ISD filter that the dynamic slope $\beta_{t|t}$ reverts to zero when the exogenous variable is excessively large. The enhanced stability of the ISD approach allows its estimated

[3] https://finance.yahoo.com/quote/MSFT/history?p=MSFT
[4] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

learning rate to substantially exceed that of the ESD model ($\hat{H} = 0.0169$ versus $\hat{H} = 0.0092$, respectively), which explains the ISD's higher sensitivity during non-crisis times.

6.2 Time-varying growth-at-risk

For policymakers, monitoring macroeconomic downside risk is crucial. One popular approach for macroeconomic risk assessment is the growth-at-risk (GaR) framework, which refers to conditional lower quantiles of GDP growth. GaR, and its relationship with financial/economic conditions, is typically estimated using quantile regressions (QRs; see Koenker and Hallock, 2001) with exogenous predictor variables (e.g., Adrian et al., 2019).

We propose to endogenously update a time-varying conditional quantile by postulating an asymmetric Laplace distribution with a time-varying location. Maximizing such a density is equivalent to minimizing Koenker and Bassett's (1978) QR check function (see Koenker and Machado, 1999). The ESD update for the τ-level quantile at time t with $0 < \tau < 1$, denoted by $q^{\text{ex}}_{t|t}(\tau)$, is given by

$$q^{\text{ex}}_{t|t}(\tau) = q_{t|t-1}(\tau) - 1[y_t < q_{t|t-1}(\tau)]\frac{H(1-\tau)}{\sigma} + 1[y_t > q_{t|t-1}(\tau)]\frac{H\tau}{\sigma}, \tag{26}$$

where y_t denotes the GDP growth rate in period t, while $1[\cdot]$ equals an indicator function equal to one if the condition in square brackets is satisfied, and zero otherwise. In addition, $H > 0$ and $\sigma > 0$ denote the scalar learning rate and a dispersion parameter, respectively, both of which are assumed to be constant over time. The ESD update (26) yields a downward adjustment of $H(1-\tau)/\sigma$ if the observed growth y_t falls below the quantile prediction $q_{t|t-1}(\tau)$ and an upward adjustment of $H\tau/\sigma$ if y_t exceeds $q_{t|t-1}(\tau)$. No adjustment is made (i.e., we set $q^{\text{ex}}_{t|t}(\tau) = q_{t|t-1}(\tau)$) if the observation exactly matches the predicted quantile (i.e., $y_t = q_{t|t-1}(\tau)$). Due to the non-differentiability of the Laplace density at this point, the explicit gradient update must be interpreted in a generalized sense, involving a subgradient. Apart from this probability-zero event, the explicit quantile update $q^{\text{ex}}_{t|t}(\tau)$ is identical to the limiting version of Engle and Manganelli's (2004) adaptive CAViaR update.

A disadvantage of the ESD update is that, due to the fixed size of the quantile adjustment, the updated quantile may overshoot the observation y_t. This issue is addressed by the ISD update, which can be computed in closed form and is denoted by $q_{t|t}(\tau)$ for the τ-level quantile at time t. The implicit update is equivalent to the explicit update as long as the latter does not overshoot the observation y_t. Otherwise, it equals y_t:

$$q_{t|t}(\tau) = \begin{cases} \min\{y_t, q^{\text{ex}}_{t|t}(\tau)\}, & y_t > q_{t|t-1}(\tau), \\ \max\{y_t, q^{\text{ex}}_{t|t}(\tau)\}, & y_t \leq q_{t|t-1}(\tau). \end{cases} \tag{27}$$

This demonstrates that in this setting, too, the ISD update is essentially a shrunken version of the ESD update. Specifically, if y_t is above (below) the predicted quantile $q_{t|t-1}(\tau)$, the ISD update follows the ESD update upward (downward), but only up to the value y_t. As a result, the implicit update is capped at y_t. This is advantageous, as overshooting (beyond y_t) decreases the model fit.

Quantile crossing poses an important practical problem when simultaneously tracking multiple quantiles using QRs. Thanks to the particular form of the update (27), the ISD model can ensure appropriate ordering of the quantiles using simple parameter restrictions. Specifically, if all quantile updates share the same learning rate H and dispersion parameter σ, the updated quantiles remain correctly ordered. To illustrate, consider an observation y_t that falls between the predictions of two different quantiles, meaning the estimate of the higher quantile must be updated downward and the estimate of the lower quantile upward. Because the ISD update is capped at the observation y_t, it is, by design, not possible for the two quantile estimates to cross. In contrast, for the ESD update, no positive learning rate exists that is guaranteed to prevent such crossings; this can be interpreted as a practical ramification of applying the ESD update to a logarithmic density without a Lipschitz-continuous gradient. To maintain the correct ordering of quantiles not only in the update but also in the prediction step, we specify the latter as

$$q_{t+1|t}(\tau) \;=\; c(\tau)\,(1-\Phi) \;+\; \Phi\, q_{t|t}(\tau) \;+\; \gamma\, x_t, \tag{28}$$

with an autoregressive parameter $\Phi \in [0, 1)$ that is common across quantiles and intercepts $c(\tau)$ that are strictly ordered in τ. Furthermore, x_t denotes an exogenous variable available at time t with common slope parameter γ.[5]

We estimate the 5, 10, 25, and 50 percent GaR with the ISD and ESD models using quarterly U.S. GDP growth rates from 1971Q1 to 2021Q4. For the exogenous variable x_t, we follow Adrian et al. (2019) in using the National Financial Conditions Index (NFCI), where quarterly values are constructed by averaging the corresponding weekly values. Both time series were obtained from the FRED database.[6] To reduce the number of parameters to estimate, we use a targeting approach (see, e.g., Engle, 2002 in the context of dynamic covariance modeling) and set $c(\tau)$ equal to the corresponding full-sample empirical quantiles. The remaining static parameters are estimated in a composite-likelihood fashion, comparable to Zou and Yuan (2008). We fix the scale parameter $\sigma = 1$, as it does not influence the

[5]While it may be useful to allow different quantiles to have different sensitivities to the exogenous input x_t, this has the potential to introduce quantile crossings. Moreover, for our application, the likelihood improvement of quantile specific slopes $\gamma(\tau)$ is too small to justify the additional model complexity.

[6]See https://fred.stlouisfed.org/series/GDP and https://fred.stlouisfed.org/series/NFCI.

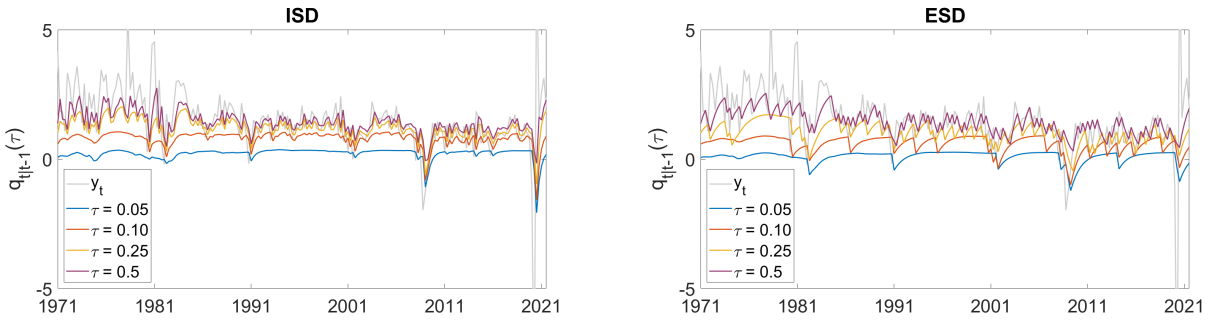


Figure 4: Growth-at-risk estimates for the ISD and ESD models for $\tau = 0.05$, $\tau = 0.10$, $\tau = 0.25$, and $\tau = 0.50$, 1971Q1 to 2021Q4.

quantile dynamics and can be treated as a nuisance parameter (e.g., Geraci and Bottai, 2007). Our postulated log-likelihood function equals the sum of four logarithmic Laplace densities, of which three are asymmetric and one (corresponding to the median) is symmetric.

Figure 4 displays the 5, 10, 25, and 50 percent GaR estimates obtained from the ISD (27) and ESD (26) models. The ISD model appears to be more responsive; for example, it shows substantially larger downward adjustments during the onset of the COVID-19 pandemic in 2020Q2 and a faster recovery after the crisis than the ESD model. This is attributable to the enhanced stability of the implicit update (27) relative to the explicit update (26), whereby the estimated learning rate H of the ISD model greatly exceeds that of the ESD model ($\hat{H} = 4.002$ and $\hat{H} = 0.804$, respectively). Furthermore, Figure 4 shows regular quantile crossings for the ESD model, while the ISD quantiles remain strictly ordered at all times. For example, the ESD update of the median ($\tau = 0.50$) frequently cuts across (i.e., overshoots) the observation y_t, and occasionally then crosses the estimate of the first quartile ($\tau = 0.25$). In line with Adrian et al. (2019), we find a negative effect of the NFCI ($\hat{\gamma} = -0.052$ and $\hat{\gamma} = -0.019$ for ISD and ESD, respectively), such that higher NFCI values reflecting tighter financial conditions correspond to more negative quantile estimates.

7 Conclusion

This article introduced a novel framework for updating time-varying parameters in an observation-driven setting. Specifically, we proposed an implicit score-driven (ISD) update that maximizes, at each point in time, the logarithmic observation density subject to a quadratic penalty centered at the one-step-ahead prediction. The name originates from the first-order condition associated with this maximization, which can be written as an implicit stochastic-gradient update. We derived model invertibility for the class of (possibly misspecified) log-concave observation densities and formulated sufficient conditions for a global

contraction of the parameter update toward a pseudo-truth at every time step. We demonstrated that the class of explicit score-driven (ESD) models—known variously as dynamic conditional score (DCS; Harvey, 2013) or generalized autoregressive score (GAS; Creal et al., 2013) models—can be obtained within the ISD framework by replacing the logarithmic observation density at every time step by its local-linear approximation around the prediction. Unlike the ESD update, the ISD update requires neither Lipschitz continuity of the gradient of the logarithmic observation density nor a small learning rate. A simulation study confirmed the theoretical advantages of the proposed method, and empirical benefits were demonstrated in two illustrations involving asset pricing and growth-at-risk.

References

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). Vulnerable growth. *American Economic Review* **109**, 1263–89.

Akyildiz, Ömer Deniz, Emilie Chouzenoux, Víctor Elvira, and Joaquín Míguez (2019). A probabilistic incremental proximal gradient method. *IEEE Signal Processing Letters* **26**, 1257–1261.

Amari, Shun-ichi (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing* **5**, 185–196.

Anderson, Brian DO and John B Moore (2012). Optimal filtering. Prentice-Hall.

Artemova, Mariia, Francisco Blasques, Janneke van Brummelen, and Siem Jan Koopman (2022a). Score-driven models: Methodology and theory. *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Artemova, Mariia, Francisco Blasques, Janneke van Brummelen, and Siem Jan Koopman (2022b). Score-driven models: Methods and applications. *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Asi, Hilal and John C Duchi (2019). Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization* **29**, 2257–2290.

Bauschke, Heinz H, Jonathan M Borwein, and Patrick L Combettes (2003). Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization* **42**, 596–636.

Bertsekas, Dimitri P (1996). Incremental least squares methods and the extended Kalman filter. *SIAM Journal on Optimization* **6**, 807–822.

Bianchi, Pascal (2016). Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization* **26**, 2235–2260.

Bierman, Gerald J (1977). Factorization methods for discrete sequential estimation. Vol. **128**. Mathematics in Science and Engineering. Academic Press.

Blasques, Francisco, Janneke van Brummelen, Siem Jan Koopman, and André Lucas (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics* **227**, 325–346.

Blasques, Francisco, Paolo Gorgi, Siem Jan Koopman, and Olivier Wintenberger (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics* **12**, 1019–1052.

Blasques, Francisco, Siem Jan Koopman, and André Lucas (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika* **102**, 325–343.

Bof, Nicoletta, Ruggero Carli, and Luca Schenato (2018). Lyapunov theory for discrete time systems. *Preprint arXiv:1809.05289*.

Bottou, Léon (2012). Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève Orr, and Klaus-Robert Müller. Springer, 421–436.

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

Cesa-Bianchi, Nicolò and Francesco Orabona (2021). Online learning algorithms. *Annual Review of Statistics and Its Application* **8**, 165–190.

Chopin, Nicolas and Omiros Papaspiliopoulos (2020). An Introduction to Sequential Monte Carlo. Springer.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics* **28**, 777–795.

Creal, Drew, Bernd Schwaab, Siem Jan Koopman, and André Lucas (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics* **96**, 898–915.

Diniz, Paulo SR (1997). Adaptive filtering. Vol. **4**. Springer.

Engle, Robert F (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* **20**, 339–350.

Engle, Robert F and Simone Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* **22**, 367–381.

Fearnhead, Paul and Loukia Meligkotsidou (2004). Exact filtering for partially observed continuous time models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66**, 771–789.

Geraci, Marco and Matteo Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**, 140–154.

Gorgi, Paolo (2020). Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1325–1347.

Grimmer, Benjamin, Haihao Lu, Pratik Worah, and Vahab Mirrokni (2023). The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming* **201**, 373–407.

Hare, Warren and Claudia Sagastizábal (2009). Computing proximal points of nonconvex functions. *Mathematical Programming* **116**, 221–258.

Harvey, Andrew (2013). Dynamic models for volatility and heavy tails: With applications to financial and economic time series. Vol. **52**. Econometric Society Monograph. New York: Cambridge University Press.

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

Harvey, Andrew and Rutger-Jan Lange (2017). Volatility modeling with a generalized t distribution. *Journal of Time Series Analysis* **38**, 175–190.

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

Jagannathan, Ravi and Zhenyu Wang (1996). The conditional CAPM and the cross-section of expected returns. *The Journal of Finance* **51**, 3–53.

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

Karimi, Hamed, Julie Nutini, and Mark Schmidt (2016). "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition". *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, 795–811.

Koenker, Roger and Gilbert Bassett (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* **46**, 33–50.

Koenker, Roger and Kevin F Hallock (2001). Quantile regression. *Journal of Economic Perspectives* **15**, 143–156.

Koenker, Roger and Jose AF Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94**, 1296–1310.

Koopman, Siem Jan, André Lucas, and Marcel Scharth (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *The Review of Economics and Statistics* **98**, 97–110.

Koyama, Shinsuke, Lucia Castellanos Pérez-Bolde, Cosma Rohilla Shalizi, and Robert E Kass (2010). Approximate methods for state-space models. *Journal of the American Statistical Association* **105**, 170–180.

Kulis, Brian and Peter L Bartlett (2010). Implicit online learning. *Proceedings of the 27th International Conference on Machine Learning*, 575–582.

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

Lange, Rutger-Jan (2024b). Short and simple introduction to Bellman filtering and smoothing. *Preprint arXiv:2405.12668*.

Opschoor, Anne, Pawel Janus, André Lucas, and Dick van Dijk (2018). New HEAVY models for fat-tailed realized covariances and returns. *Journal of Business & Economic Statistics* **36**, 643–657.

Orabona, Francesco (2019). A modern introduction to online learning. *Preprint arXiv:1912.13213*.

Parikh, Neal and Stephen Boyd (2014). Proximal algorithms. *Foundations and Trends® in Optimization* **1**, 127–239.

Patrascu, Andrei and Ion Necoara (2018). Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *The Journal of Machine Learning Research* **18**, 7204–7245.

Polson, Nicholas G., James G. Scott, and Brandon T. Willard (2015). Proximal Algorithms in Statistics and Machine Learning. *Statistical Science* **30**, 559–581.

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

Rockafellar, R Tyrrell (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14**, 877–898.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **71**, 319–392.

Ryu, Ernest K and Stephen Boyd (2016). Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. *Author website*.

Simonetto, Andrea and Paolo Massioni (2024). Nonlinear optimization filters for stochastic time-varying convex optimization. *International Journal of Robust and Nonlinear Control* **34**, 8065–8089.

Stock, James H and Mark W Watson (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics* **14**, 11–30.

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

Teräsvirta, Timo (2009). An introduction to univariate GARCH models. *Handbook of Financial Time Series*. Ed. by Torben Gustav Andersen, Richard A Davis, Jens-Peter Kreiß, and Thomas V Mikosch. Springer, 17–42.

Toulis, Panos and Edoardo M Airoldi (2015). Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Statistics and Computing* **25**, 781–795.

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

Toulis, Panos, Thibaut Horel, and Edoardo M Airoldi (2021). The proximal Robbins-Monro method. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83**, 188–212.

Toulis, Panos, Dustin Tran, and Edoardo M Airoldi (2016). Towards stability and optimality in stochastic gradient descent. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* **51**, 1290–1298.

Zou, Hui and Ming Yuan (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36**, 1108–1126.

Appendix to "Implicit score-driven filters for time-varying parameter models"

A Proofs

A.1 Proposition 1: Relationship ISD and ESD updates

By Assumption 2 we have that the regularized log likelihood $f(\theta|y_t, \theta_{t|t-1})$ is concave in θ with probability one in y_t. As a result, we have for almost every y_t that

$$f(\theta_{t|t}|y_t, \theta_{t|t-1}) \leq f(\theta_{t|t-1}|y_t, \theta_{t|t-1}) + \langle \theta_{t|t} - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1}) \rangle, \tag{A.1}$$

reordering and using the fact that $\theta_{t|t}$ maximizes $f(\theta|y_t, \theta_{t|t-1})$ we obtain the desired result:

$$\langle \theta_{t|t} - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1}) \rangle \geq f(\theta_{t|t}|y_t, \theta_{t|t-1}) - f(\theta_{t|t-1}|y_t, \theta_{t|t-1}) \geq 0. \tag{A.2}$$

In the scalar case, the first-order condition (FOC) and strict positivity of the learning rate imply that $\nabla(y_t|\theta_{t|t-1})\nabla(y_t|\theta_{t|t}) \geq 0$. Furthermore, $\nabla(y_t|\theta_{t|t}) = 0$ produces $\theta_{t|t} = \theta_{t|t-1}$, in turn implying that $\nabla(y_t|\theta_{t|t-1}) = \nabla(y_t|\theta_{t|t}) = 0$. Conversely, if $\nabla(y_t|\theta_{t|t-1}) = 0$, we have that $\theta_{t|t} = \theta_{t|t-1}$, as filling in $\theta_{t|t-1}$ solves the FOC (and Assumption 2 implies uniqueness of $\theta_{t|t}$). Therefore, $\nabla(y_t|\theta_{t|t-1}) = 0$ if and only if $\nabla(y_t|\theta_{t|t}) = 0$. Combining this with $\nabla(y_t|\theta_{t|t-1})\nabla(y_t|\theta_{t|t}) \geq 0$, we obtain $\text{sgn}(\nabla(y_t|\theta_{t|t})) = \text{sgn}(\nabla(y_t|\theta_{t|t-1}))$, which is the score-equivalence as defined in Blasques et al. (2015).

Next, we use second-order differentiability (Assumption 4b) to write the ISD update as a curvature-corrected ESD update. Specifically, we have that

$$\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1}) = -\mathcal{I}_{t|t}(\theta_{t|t} - \theta_{t|t-1}), \tag{A.3}$$

where $\mathcal{I}_{t|t}$ is the negative average Hessian between $\theta_{t|t-1}$ and $\theta_{t|t}$:

$$\mathcal{I}_{t|t} := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t-1} + (1-u)\,\theta_{t|t}} \mathrm{d}u. \tag{A.4}$$

Roughly put, relationship (A.3) can be viewed as a multivariate analog of the mean-value theorem. The integral form is necessary because, unlike the scalar case, there need not be any single point θ for which the Hessian is exactly equal to $-\mathcal{I}_{t|t}$. Using the ISD FOC for $\theta_{t|t} - \theta_{t|t-1}$ on the right-hand side of (A.3) and rearranging yields:

$$(I_K + \mathcal{I}_{t|t}P_t^{-1})\nabla(y_t|\theta_{t|t}) = \nabla(y_t|\theta_{t|t-1}). \tag{A.5}$$

Next, we note that $P_t + \mathcal{I}_{t|t}$ is positive definite, which follows from strict concavity of the regularized log-likelihood (Assumption 2). That is, this assumption implies that the penalty matrix P_t strictly exceeds (in an eigenvalue sense) the Hessian of $\log p(y_t|\theta)$ for any θ with probability one in y_t. From the definition of $\mathcal{I}_{t|t}$, it follows that $P_t \succ -\mathcal{I}_{t|t} \Rightarrow P_t + \mathcal{I}_{t|t} \succ O_K$. Therefore $P_t + \mathcal{I}_{t|t}$ is invertible; premultiplying with $(P_t + \mathcal{I}_{t|t})^{-1}$ gives:

$$P_t^{-1}\nabla(y_t|\theta_{t|t}) = (P_t + \mathcal{I}_{t|t})^{-1}\nabla(y_t|\theta_{t|t-1}). \tag{A.6}$$

Using (A.6) together with the ISD FOC produces the final result,

$$\theta_{t|t} = \theta_{t|t-1} + (P_t + \mathcal{I}_{t|t})^{-1}\nabla(y_t|\theta_{t|t-1}). \tag{A.7}$$

A.2 Proposition 2: Step-size shrinkage

Using the FOCs, the difference between the implicit and explicit updates is given by:

$$\theta_{t|t}^{\text{ex}} - \theta_{t|t} = \theta_{t|t-1} + P_t^{-1}\nabla(y_t|\theta_{t|t-1}) - \theta_{t|t-1} - P_t^{-1}\nabla(y_t|\theta_{t|t}), \tag{A.8}$$

whereby rearranging and using the definition of $\mathcal{I}_{t|t}$ from (A.4) yields

$$\theta_{t|t}^{\text{ex}} - \theta_{t|t-1} = \theta_{t|t} - \theta_{t|t-1} - P_t^{-1}[\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1})] \tag{A.9}$$

$$= (I_K + P_t^{-1}\mathcal{I}_{t|t})(\theta_{t|t} - \theta_{t|t-1}), \tag{A.10}$$

Pre-multiplying with $P_t^{1/2}$, which denotes the symmetric square root of P_t, and taking the quadratic norm yields

$$\|\theta^{\text{ex}}_{t|t} - \theta_{t|t-1}\|^2_{P_t} = \|\theta_{t|t} - \theta_{t|t-1}\|^2_{(I_K + P_t^{-1}\mathcal{I}_{t|t})' P_t (I_K + P_t^{-1}\mathcal{I}_{t|t})} \tag{A.11}$$

$$= \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t + 2\mathcal{I}_{t|t} + \mathcal{I}_{t|t} P_t^{-1} \mathcal{I}_{t|t}} \tag{A.12}$$

$$= \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t} + 2\|\theta_{t|t} - \theta_{t|t-1}\|^2_{\mathcal{I}_{t|t}} + \|\mathcal{I}_{t|t}(\theta_{t|t} - \theta_{t|t-1})\|^2_{P_t^{-1}} \tag{A.13}$$

$$\geq \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t} + 2\lambda_{\min}(\mathcal{I}_{t|t})\|\theta_{t|t} - \theta_{t|t-1}\|^2 + \lambda_{\min}(P_t^{-1})\|\theta_{t|t} - \theta_{t|t-1}\|^2_{\mathcal{I}^2_{t|t}} \tag{A.14}$$

$$\geq \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t} + 2\alpha_t\|\theta_{t|t} - \theta_{t|t-1}\|^2 + \frac{\lambda_{\min}(\mathcal{I}^2_{t|t})}{\lambda_{\max}(P_t)}\|\theta_{t|t} - \theta_{t|t-1}\|^2 \tag{A.15}$$

$$\geq \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t} + \left(2\alpha_t + \frac{\alpha_t^2}{\lambda_{\max}(P_t)}\right)\|\theta_{t|t} - \theta_{t|t-1}\|^2 \tag{A.16}$$

$$\geq \left(1 + \frac{2\alpha_t}{\lambda_{\max}(P_t)} + \frac{\alpha_t^2}{\lambda_{\max}(P_t)^2}\right)\|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t} = \left(\frac{\lambda_{\max}(P_t) + \alpha_t}{\lambda_{\max}(P_t)}\right)^2 \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t}. \tag{A.17}$$

Here the fifth line uses $\lambda_{\min}(\mathcal{I}_{t|t}) \geq \alpha_t$ which follows from concavity (Assumption 5) and the definition of $\mathcal{I}_{t|t}$ in (A.4) as the average negative hessian. Because $\alpha_t \geq 0$ it also follows that $\lambda_{\min}(\mathcal{I}^2_{t|t}) \geq \alpha_t^2$, which is used in the sixth line. The final line then uses $\|x\|^2_{P_t} \leq \lambda_{\max}(P_t)\|x\|^2$, which implies that $\|x\|^2 \geq \frac{1}{\lambda_{\max}(P_t)}\|x\|^2_{P_t}$ for arbitrary $K \times 1$ vector x and using that $\lambda_{\max}(P_t) > 0$. Note that $2\alpha_t + \frac{\alpha_t^2}{\lambda_{\max}(P_t)} \geq 0$, such that the correct sign is maintained. Finally, rearranging the last expression above gives the desired result:

$$\left\|\theta_{t|t} - \theta_{t|t-1}\right\|^2_{P_t} \leq \left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2 \left\|\theta^{\text{ex}}_{t|t} - \theta_{t|t-1}\right\|^2_{P_t}. \tag{A.18}$$

A.3 Lemma 1: Prediction-to-update stability

Consider two predictions $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$ that are updated based on the observation y_t to $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$, respectively. For the ISD update, we may write

$$\theta_{t|t} - \tilde{\theta}_{t|t} = \theta_{t|t-1} - \tilde{\theta}_{t|t-1} + P_t^{-1}[\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\tilde{\theta}_{t|t})] \tag{A.19}$$

$$= \theta_{t|t-1} - \tilde{\theta}_{t|t-1} - P_t^{-1}\tilde{\mathcal{I}}_{t|t}(\theta_{t|t} - \tilde{\theta}_{t|t}), \tag{A.20}$$

where $\tilde{\mathcal{I}}_{t|t}$ is the negative average Hessian between $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$:

$$\tilde{\mathcal{I}}_{t|t} := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t} + (1-u)\,\tilde{\theta}_{t|t}} \mathrm{d}u. \tag{A.21}$$

Reordering of (A.20) yields

$$\theta_{t|t-1} - \tilde{\theta}_{t|t-1} = (I_K + P_t^{-1}\tilde{\mathcal{I}}_{t|t})(\theta_{t|t} - \tilde{\theta}_{t|t}). \tag{A.22}$$

Next, pre-multiplying with $P_t^{1/2}$ and taking the quadratic norm gives

$$\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2 = \|\theta_{t|t} - \tilde{\theta}_{t|t}\|_{P_t+2\tilde{\mathcal{I}}_{t|t}+\tilde{\mathcal{I}}_{t|t}P_t^{-1}\tilde{\mathcal{I}}_{t|t}}^2. \tag{A.23}$$

Using the same steps as in (A.11)-(A.18), we obtain

$$\left\|\theta_{t|t} - \tilde{\theta}_{t|t}\right\|_{P_t}^2 \leq \left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2 \left\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\right\|_{P_t}^2. \tag{A.24}$$

For the ESD update, we may write

$$\begin{aligned}
\theta_{t|t}^{\text{ex}} - \tilde{\theta}_{t|t}^{\text{ex}} &= \theta_{t|t-1} - \tilde{\theta}_{t|t-1} + P_t^{-1}[\nabla(y_t|\theta_{t|t-1}) - \nabla(y_t|\tilde{\theta}_{t|t-1})] && \text{(A.25)}\\
&= (I_K - P_t^{-1}\tilde{\mathcal{I}}_{t|t-1})(\theta_{t|t-1} - \tilde{\theta}_{t|t-1}), && \text{(A.26)}
\end{aligned}$$

where $\tilde{\mathcal{I}}_{t|t-1}$ is the negative average Hessian between $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$:

$$\tilde{\mathcal{I}}_{t|t-1} := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t-1} + (1-u)\,\tilde{\theta}_{t|t-1}} \mathrm{d}u. \tag{A.27}$$

Next, pre-multiplying with $P_t^{1/2}$ and taking the quadratic norm gives

$$\begin{aligned}
\|\theta_{t|t}^{\text{ex}} - \tilde{\theta}_{t|t}^{\text{ex}}\|_{P_t}^2 &= \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{(I_K - P_t^{-1}\tilde{\mathcal{I}}_{t|t-1})' P_t (I_K - P_t^{-1}\tilde{\mathcal{I}}_{t|t-1})}^2 && \text{(A.28)}\\
&= \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t - 2\tilde{\mathcal{I}}_{t|t-1} + \tilde{\mathcal{I}}_{t|t-1}P_t^{-1}\tilde{\mathcal{I}}_{t|t-1}}^2 && \text{(A.29)}\\
&= \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2 + \|\tilde{\mathcal{I}}_{t|t-1}^{1/2}(\theta_{t|t-1} - \tilde{\theta}_{t|t-1})\|_{\tilde{\mathcal{I}}_{t|t-1}^{1/2} P_t^{-1} \tilde{\mathcal{I}}_{t|t-1}^{1/2} - 2I_K}^2 && \text{(A.30)}\\
&\leq \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2 + \lambda_{\max}(\tilde{\mathcal{I}}_{t|t-1}^{1/2} P_t^{-1} \tilde{\mathcal{I}}_{t|t-1}^{1/2} - 2I_K)\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\tilde{\mathcal{I}}_{t|t-1}}^2, && \text{(A.31)}
\end{aligned}$$

where $\lambda_{\max}(\tilde{\mathcal{I}}_{t|t-1}^{1/2} P_t^{-1} \tilde{\mathcal{I}}_{t|t-1}^{1/2} - 2I_K) = \lambda_{\max}(\tilde{\mathcal{I}}_{t|t-1}P_t^{-1}) - 2 \leq \lambda_{\max}(\tilde{\mathcal{I}}_{t|t-1})\lambda_{\max}(P_t^{-1}) - 2 \leq L_t/\lambda_{\min}(P_t) - 2 \leq 0$ and the final argument follows from the assumption $\lambda_{\min}(P_t) \geq L_t/2$. This means that the final term in (A.31) is negative. Continuing, we obtain

$$\begin{aligned}
\|\theta_{t|t}^{\text{ex}} - \tilde{\theta}_{t|t}^{\text{ex}}\|_{P_t}^2 &\leq \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2 - [2 - L_t/\lambda_{\min}(P_t)]\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\tilde{\mathcal{I}}_{t|t-1}}^2 && \text{(A.32)}\\
&\leq \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2 - \alpha_t[2 - L_t/\lambda_{\min}(P_t)]\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2 && \text{(A.33)}
\end{aligned}$$

$$\leq \left(1 - \frac{\alpha_t[2 - L_t/\lambda_{\min}(P_t)]}{\lambda_{\max}(P_t)}\right) \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{P_t} \tag{A.34}$$

$$= \frac{\lambda_{\max}(P_t) - \alpha_t[2 - L_t/\lambda_{\min}(P_t)]}{\lambda_{\max}(P_t)} \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{P_t}, \tag{A.35}$$

where the second line uses $\lambda_{\min}(\tilde{\mathcal{I}}_{t|t-1}) \geq \alpha_t \geq 0$ by concavity (Assumption 5) and the third line uses that $-\|x\|^2 \leq -\frac{1}{\lambda_{\max}(P_t)}\|x\|^2_{P_t}$ for arbitrary $K \times 1$ vector x. Finally, we note that $\lambda_{\max}(P_t) - \alpha_t\left[2 - L_t/\lambda_{\min}(P_t)\right] \geq \lambda_{\max}(P_t) - \alpha_t\left[2 - \alpha_t/\lambda_{\max}(P_t)\right] = \frac{1}{\lambda_{\max}(P_t)}(\lambda_{\max}(P_t)^2 - 2\alpha_t\lambda_{\max}(P_t) + \alpha_t^2) = \frac{1}{\lambda_{\max}(P_t)}(\lambda_{\max}(P_t) - \alpha_t)^2 \geq 0$, such that contraction coefficient is indeed contained in the unit interval:

$$\|\theta^{\text{ex}}_{t|t} - \tilde{\theta}^{\text{ex}}_{t|t}\|^2_{P_t} \leq \underbrace{\frac{\lambda_{\max}(P_t) - \alpha_t[2 - L_t/\lambda_{\min}(P_t)]}{\lambda_{\max}(P_t)}}_{\in [0,1], \text{ contraction coefficient}} \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{P_t}. \tag{A.36}$$

A.4 Lemma 2: Prediction-to-prediction stability

The update-to-prediction mapping from time t to $t+1$ can be written as

$$\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|^2_{P_t} = \|\Phi(\theta_{t|t} - \tilde{\theta}_{t|t})\|^2_{P_t} = -\|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t - \Phi' P_t \Phi} + \|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t} \tag{A.37}$$

$$\leq -\lambda_{\min}(P_t - \Phi' P_t \Phi)\|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2 + \|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t} \tag{A.38}$$

$$\leq \varepsilon_{1,t}\|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t}, \tag{A.39}$$

where the second line uses that $\lambda_{\min}(P_t - \Phi' P_t \Phi) \geq 0$ by positive semi-definiteness of $P_t - \Phi' P_t \Phi$, while the last line uses $-\|\cdot\|^2 \leq -\lambda_{\max}(P_t)^{-1}\|\cdot\|^2_{P_t}$. Here $\varepsilon_{1,t}$ is given by

$$\varepsilon_{1,t} = \frac{\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)}{\lambda_{\max}(P_t)}. \tag{A.40}$$

By positive definiteness of P_t it follows that $\Phi' P_t \Phi$ is positive semi-definite due to its quadratic form. Therefore, we have that $0 \leq \lambda_{\max}(\Phi' P_t \Phi) = \lambda_{\max}(P_t - (P_t - \Phi' P_t \Phi)) \leq \lambda_{\max}(P_t) + \lambda_{\max}(-(P_t - \Phi' P_t \Phi)) = \lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi) \leq \lambda_{\max}(P_t)$, such that $\varepsilon_{1,t} \in [0,1]$. If $P_t - \Phi' P_t \Phi$ is positive definite, we have that $\varepsilon_{1,t} \in [0,1)$.

In addition, from Lemma 1, we have

$$\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t} \leq \varepsilon_{2,t}\|\tilde{\theta}_{t|t-1} - \theta_{t|t-1}\|^2_{P_t}, \quad \varepsilon_{2,t} = \left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2, \tag{A.41}$$

where $\varepsilon_{2,t} \in [0,1]$ if $\alpha_t \geq 0$ and $\varepsilon_{2,t} \in [0,1)$ if $\alpha_t > 0$. Combining (A.39) and (A.41), we

obtain

$$\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|_{P_t}^2 \leq \kappa_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2, \tag{A.42}$$

where κ_t is given as

$$\kappa_t = \varepsilon_{1,t}\varepsilon_{2,t} = \frac{\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)}{\lambda_{\max}(P_t)} \frac{\lambda_{\max}(P_t)^2}{(\lambda_{\max}(P_t) + \alpha_t)^2} \tag{A.43}$$

$$= \frac{\lambda_{\max}(P_t)[\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)]}{(\lambda_{\max}(P_t) + \alpha_t)^2}. \tag{A.44}$$

If either $\alpha_t > 0$ or $P_t - \Phi' P_t \Phi \succ O_K$ we obtain $\kappa_t \in [0, 1)$, which concludes the proof.

A.5 Theorem 1: Invertibility

By assumption there exists a $\bar{P}$ such that we have for all P_t that $\kappa_t P_t \prec \rho_t \bar{P} \preceq P_t$ for some $\rho_t > 0$. This condition implies that the prediction-to-prediction mapping from time t to $t+1$ is strictly contracting in the norm $\|\cdot\|_{\rho_t \bar{P}}$. To see this, we may write

$$\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|_{\rho_t \bar{P}}^2 \leq \|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|_{P_t}^2 \leq \kappa_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{P_t}^2 \tag{A.45}$$

$$= -\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\rho_t \bar{P} - \kappa_t P_t}^2 + \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\rho_t \bar{P}}^2 \tag{A.46}$$

$$\leq \delta_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\rho_t \bar{P}}^2, \tag{A.47}$$

where δ_t is given as

$$\delta_t = \frac{\lambda_{\max}(\rho_t \bar{P}) - \lambda_{\min}(\rho_t \bar{P} - \kappa_t P_t)}{\lambda_{\max}(\rho_t \bar{P})}. \tag{A.48}$$

Due to the condition $\rho_t \bar{P} - \kappa_t P_t \succeq \rho_t A \succ 0$, we obtain that $\delta_t \in [0, \delta]$, where δ is given as

$$\delta = \frac{\lambda_{\max}(\rho_t \bar{P}) - \lambda_{\min}(\rho_t A)}{\lambda_{\max}(\rho_t \bar{P})} = \frac{\lambda_{\max}(\bar{P}) - \lambda_{\min}(A)}{\lambda_{\max}(\bar{P})}, \tag{A.49}$$

where due to positive definiteness of $\bar{P}$ and A we have that $\delta \in (0, 1)$. It now follows that

$$\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|_{\bar{P}}^2 \leq \delta \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\bar{P}}^2, \tag{A.50}$$

such that every prediction-to-prediction mapping is strictly contracting in a common norm $\|\cdot\|_{\bar{P}}^2$ with at least strength of contraction $\delta \in (0, 1)$. Therefore, for any $c \in (1, \frac{1}{\delta})$, we have

$$\lim_{t \to \infty} c^t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{\bar{P}}^2 \to 0, \tag{A.51}$$

By norm equivalence it follows that this difference convergences to 0 in any norm.

A.6 Theorem 2: Contraction to the NDR

We write the first-order condition of the ISD update as follows

$$P_t^{1/2}(\theta_{t|t} - \theta_{t|t-1}) = P_t^{-1/2}\nabla(y_t|\theta_{t|t}), \tag{A.52}$$

where adding $P_t^{-1/2}\nabla(y_t|\theta_t^\star) - P_t^{1/2}\theta_t^\star$ to both sides and rearranging gives

$$P_t^{1/2}(\theta_{t|t} - \theta_t^\star) + P_t^{-1/2}(\nabla(y_t|\theta_t^\star) - \nabla(y_t|\theta_{t|t})) = P_t^{1/2}(\theta_{t|t-1} - \theta_t^\star) + P_t^{-1/2}\nabla(y_t|\theta_t^\star). \tag{A.53}$$

Next, we write the difference in gradients, $\nabla(y_t|\theta_t^\star) - \nabla(y_t|\theta_{t|t})$, as the product of the negative average Hessian and the difference in two points. That is,

$$(P_t^{1/2} + P_t^{-1/2}\mathcal{I}_{t|t}^\star)(\theta_{t|t} - \theta_t^\star) = P_t^{1/2}(\theta_{t|t-1} - \theta_t^\star) + P_t^{-1/2}\nabla(y_t|\theta_t^\star), \tag{A.54}$$

where $\mathcal{I}_{t|t}^\star \succeq \alpha_t I_K \succeq O_K$ is the negative average $K \times K$ Hessian between $\theta_{t|t}$ and $\theta_t^\star$:

$$\mathcal{I}_{t|t}^\star := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t} + (1-u)\,\theta_t^\star} \mathrm{d}u. \tag{A.55}$$

Next, we consider the quadratic norm of (A.54) and take the expectation over y_t with respect to the DGP. Reordering and using that $\mathop{\mathbb{E}}_{y_t}[\nabla(y_t|\theta_t^\star)] = 0$ by Assumption 6 gives:

$$\underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t} - \theta_t^\star\right\|_{P_t}^2\right]}_{\text{MSE after update}} = \underbrace{\left\|\theta_{t|t-1} - \theta_t^\star\right\|_{P_t}^2}_{\text{SE before update}} - \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t} - \theta_t^\star\right\|_{2\mathcal{I}_{t|t}^\star + \mathcal{I}_{t|t}^\star P_t^{-1}\mathcal{I}_{t|t}^\star}^2\right]}_{\geq 0,\ \text{contractive force}} + \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|_{P_t^{-1}}^2\right]}_{\geq 0,\ \text{expansive force}}. \tag{A.56}$$

The positivity of the contractive force is apparent from the positive semi-definiteness of $\mathcal{I}_{t|t}^\star$, which implies that also $2\mathcal{I}_{t|t}^\star + \mathcal{I}_{t|t}^\star P_t^{-1}\mathcal{I}_{t|t}^\star$ is positive semi-definite.

For the second result, we write the ESD update as follows

$$P_t^{1/2}(\theta_{t|t}^{\text{ex}} - \theta_{t|t-1}) = P_t^{-1/2}\nabla(y_t|\theta_{t|t-1}), \tag{A.57}$$

where subtracting $P_t^{1/2}\theta_t^\star$ on both sides and taking the quadratic norm yields:

$$\|\theta_{t|t}^{\text{ex}} - \theta_t^\star\|_{P_t}^2 = \|\theta_{t|t-1} - \theta_t^\star\|_{P_t}^2 + 2\langle\theta_{t|t-1} - \theta_t^\star, \nabla(y_t|\theta_{t|t-1})\rangle + \|\nabla(y_t|\theta_{t|t-1})\|_{P_t^{-1}}^2. \tag{A.58}$$

We now take the expectation over y_t with respect to the DGP. Using that $\mathop{\mathbb{E}}_{y_t}[\nabla(y_t|\theta_t^\star)] = 0$ by Assumption 6, we also subtract $2\mathop{\mathbb{E}}_{y_t}[\langle\theta_{t|t-1} - \theta_t^\star, \nabla(y_t|\theta_t^\star)\rangle] = 0$ from the right-hand side

and write the difference in gradients, $\nabla(y_t|\theta_{t|t-1}) - \nabla(y_t|\theta_t^\star)$, as the product of the negative average Hessian and the difference in two points. This gives

$$\underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t}^{\text{ex}} - \theta_t^\star\right\|_{P_t}^2\right]}_{\text{MSE after update}} = \underbrace{\left\|\theta_{t|t-1} - \theta_t^\star\right\|_{P_t}^2}_{\text{SE before update}} - \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t-1} - \theta_t^\star\right\|_{2\mathcal{I}_{t|t-1}^\star}^2\right]}_{\geq 0,\text{ contractive force}} + \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_{t|t-1})\right\|_{P_t^{-1}}^2\right]}_{\geq 0,\text{ expansive force}}. \quad \text{(A.59)}$$

where the positivity of the contractive force follows from the positive semi-definiteness of $\mathcal{I}_{t|t-1}^\star \succeq \alpha_t I_K \succeq O_K$, which is the negative average $K \times K$ Hessian between $\theta_{t|t-1}$ and $\theta_t^\star$:

$$\mathcal{I}_{t|t-1}^\star := -\int_0^1 \left.\frac{\partial^2 \log p(y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta = u\,\theta_{t|t-1} + (1-u)\,\theta_t^\star} \mathrm{d}u. \quad \text{(A.60)}$$

A.7 Corollary 1: Geometric contraction to the NDR

Using α_t-strong concavity and the same steps as in equations (A.11)-(A.18), we may obtain

$$\|\theta_{t|t} - \theta_t^\star\|_{P_t}^2 \leq \left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2 \|\theta_{t|t} - \theta_t^\star\|_{P_t + 2\mathcal{I}_{t|t}^\star + \mathcal{I}_{t|t}^\star P_t^{-1} \mathcal{I}_{t|t}^\star}^2. \quad \text{(A.61)}$$

Taking the expectation over y_t with respect to the DGP and combining this with (A.56) produces the final result:

$$\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t} - \theta_t^\star\right\|_{P_t}^2\right] \leq \underbrace{\left(\frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + \alpha_t}\right)^2}_{\in [0,1),\text{ contraction coefficient}} \left(\left\|\theta_{t|t-1} - \theta_t^\star\right\|_{P_t}^2 + \mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|_{P_t^{-1}}^2\right]\right). \quad \text{(A.62)}$$

B Further results

B.1 Noise-dominated region for the ESD update

To obtain also a noise-dominated region contraction for the ESD update, we may write $\nabla(y_t|\theta_{t|t-1}) = \nabla(y_t|\theta_t^\star) - \mathcal{I}_{t|t-1}^\star(\theta_{t|t-1} - \theta_t^\star)$ and substitute this in the final term. Expanding this term yields a cross-product that is hard to assess. However, we can use the general albeit loose bound $\|a+b\|_{P_t^{-1}}^2 \leq 2\|a\|_{P_t^{-1}}^2 + 2\|b\|_{P_t^{-1}}^2$ for arbitrary $K \times 1$ vectors a, b. We note that more generally, one could apply Young's inequality to the cross-term $2\langle P_t^{-1}a, b\rangle$ and fine-tune the exponent choice; the above is essentially a special case with exponent 2. Using $a = \nabla(y_t|\theta_t^\star)$ and $b = -\mathcal{I}_{t|t-1}^\star(\theta_{t|t-1} - \theta_t^\star)$, we obtain

$$\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_{t|t-1})\right\|_{P_t^{-1}}^2\right] = \mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star) - \mathcal{I}_{t|t-1}^\star(\theta_{t|t-1} - \theta_t^\star)\right\|_{P_t^{-1}}^2\right] \quad \text{(B.63)}$$

$$\leq 2\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|^2_{P_t^{-1}}\right] + 2\mathop{\mathbb{E}}_{y_t}\left[\left\|\mathcal{I}^\star_{t|t-1}(\theta_{t|t-1}-\theta_t^\star)\right\|^2_{P_t^{-1}}\right] \tag{B.64}$$

$$= 2\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|^2_{P_t^{-1}}\right] + 2\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t-1}-\theta_t^\star\right\|^2_{\mathcal{I}^\star_{t|t-1}P_t^{-1}\mathcal{I}^\star_{t|t-1}}\right]. \tag{B.65}$$

Combining this with (A.59) gives:

$$\underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta^{\text{ex}}_{t|t}-\theta_t^\star\right\|^2_{P_t}\right]}_{\text{MSE after update}} \leq \underbrace{\left\|\theta_{t|t-1}-\theta_t^\star\right\|^2_{P_t}}_{\text{SE before update}} - \underbrace{2\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t-1}-\theta_t^\star\right\|^2_{\mathcal{I}^\star_{t|t-1}-\mathcal{I}^\star_{t|t-1}P_t^{-1}\mathcal{I}^\star_{t|t-1}}\right]}_{\geq 0,\text{ contractive force}} \tag{B.66}$$

$$+ \underbrace{2\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|^2_{P_t^{-1}}\right]}_{\geq 0,\text{ expansive force}}, \tag{B.67}$$

where positivity of the new contractive force can be guaranteed using L_t-Lipschitz continuity of the gradient combined with $\lambda_{\min}(P_t) \geq L_t \Rightarrow 1 - L_t/\lambda_{\min}(P_t) \geq 0$. That is,

$$\mathop{\mathbb{E}}_{y_t}\left[\|\theta_{t|t-1}-\theta_t^\star\|^2_{\mathcal{I}^\star_{t|t-1}-\mathcal{I}^\star_{t|t-1}P_t^{-1}\mathcal{I}^\star_{t|t-1}}\right] \tag{B.68}$$

$$= \mathop{\mathbb{E}}_{y_t}\left[\|(\mathcal{I}^\star_{t|t-1})^{1/2}(\theta_{t|t-1}-\theta_t^\star)\|^2_{I_K-(\mathcal{I}^\star_{t|t-1})^{1/2}P_t^{-1}(\mathcal{I}^\star_{t|t-1})^{1/2}}\right] \tag{B.69}$$

$$\geq \lambda_{\min}(I_K-(\mathcal{I}^\star_{t|t-1})^{1/2}P_t^{-1}(\mathcal{I}^\star_{t|t-1})^{1/2})\mathop{\mathbb{E}}_{y_t}\left[\|(\mathcal{I}^\star_{t|t-1})^{1/2}(\theta_{t|t-1}-\theta_t^\star)\|^2\right] \tag{B.70}$$

$$\geq [1-\lambda_{\max}(\mathcal{I}^\star_{t|t-1}P_t^{-1})]\mathop{\mathbb{E}}_{y_t}\left[\|(\mathcal{I}^\star_{t|t-1})^{1/2}(\theta_{t|t-1}-\theta_t^\star)\|^2\right] \tag{B.71}$$

$$\geq [1-\lambda_{\max}(\mathcal{I}^\star_{t|t-1})\lambda_{\max}(P_t^{-1})]\mathop{\mathbb{E}}_{y_t}\left[\|(\mathcal{I}^\star_{t|t-1})^{1/2}(\theta_{t|t-1}-\theta_t^\star)\|^2\right] \tag{B.72}$$

$$\geq [1-L_t/\lambda_{\min}(P_t)]\mathop{\mathbb{E}}_{y_t}\left[\|(\mathcal{I}^\star_{t|t-1})^{1/2}(\theta_{t|t-1}-\theta_t^\star)\|^2\right] \geq 0. \tag{B.73}$$

B.2 Dynamic linear regression

Consider the linear regression model with dependent variable $y_t \in \mathbb{R}$ and independent variable $x_t \in \mathbb{R}^K$, that is,

$$y_t \;=\; \beta_t' x_t \;+\; \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma^2), \tag{B.74}$$

where β_t is a $K \times 1$ vector of time-varying parameters and ε_t is an i.i.d. normally distributed innovation with variance σ^2.

The log-likelihood contribution $\log p(y_t|\beta)$ is obviously twice continuously differentiable with respect to β for all y_t, such that Assumption 4b (differentiability) holds. In addition, the Hessian is equal to $-\frac{1}{\sigma^2}x_t x_t'$ and is therefore negative semi-definite. Combined with strong concavity of the penalty this means that the regularized log likelihood $f(\beta|y_t,\beta_{t|t-1},P_t) := \log p(y_t|\beta) - \frac{1}{2}\left\|\beta-\beta_{t|t-1}\right\|^2_{P_t}$ is strongly concave in β. Because $f(\beta|y_t,\beta_{t|t-1},P_t)$ is finite-valued

for any $\beta \in \mathbb{R}^K$, we have that it is thus strictly proper concave such that Assumption 2 (strictly concave regularized log likelihood) holds.

The first-order condition (FOC) of the ISD update at time t associated with the model (B.74) takes the following form

$$\beta_{t|t} = \beta_{t|t-1} + H_t \nabla(y_t|\beta_{t|t}, x_t), \tag{B.75}$$

where $H_t = P_t^{-1}$ is the learning-rate matrix and $\nabla(y_t|\beta_{t|t}, x_t)$ denotes the implicit score given as,

$$\nabla(y_t|\beta_{t|t}, x_t) \;=\; \frac{y_t - \beta'_{t|t}\, x_t}{\sigma^2}\, x_t. \tag{B.76}$$

Note that strong concavity of $f(\beta|y_t, \beta_{t|t-1}, P_t)$ and the unrestricted nature of the optimization (i.e. we maximize over $\mathbb{R}^K$) imply that if the FOC (B.75) has a solution then it is the unique global maximizer. Solving the FOC will thus also directly verify Assumptions 1 (existence) and 3 (interior solution).

Collecting all terms containing $\beta_{t|t}$ on the left-hand side, we may write the FOC in (B.75) as

$$(I_K + H_t \frac{x_t x_t'}{\sigma^2})\beta_{t|t} = \beta_{t|t-1} + H_t \frac{y_t x_t}{\sigma^2}. \tag{B.77}$$

Now using the Sherman-Morrison identity, we left-multiply with $(I_K + H_t \frac{x_t x_t'}{\sigma^2})^{-1} = I_K - \frac{H_t x_t x_t'}{\sigma^2 + x_t' H_t x_t}$, which yields

$$\beta_{t|t} = \left(I_K - \frac{H_t x_t x_t'}{\sigma^2 + x_t' H_t x_t}\right)\left(\beta_{t|t-1} + H_t \frac{y_t x_t}{\sigma^2}\right). \tag{B.78}$$

Eliminating brackets and using the notation $\|x_t\|^2_{H_t} := x_t' H_t x_t$ then gives

$$\beta_{t|t} = \beta_{t|t-1} + H_t \frac{y_t x_t}{\sigma^2} - \frac{H_t x_t x_t'}{\sigma^2 + \|x_t\|^2_{H_t}}\beta_{t|t-1} - \frac{H_t x_t x_t'}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \frac{y_t x_t}{\sigma^2}, \tag{B.79}$$

where changing the ordering using the fact that y_t, σ^2, $x_t'\beta_{t|t-1}$ and $\|x_t\|^2_{H_t}$ are scalars and again using the definition of $\|x_t\|^2_{H_t}$, we get

$$\beta_{t|t} = \beta_{t|t-1} + H_t \frac{y_t}{\sigma^2} x_t - \frac{1}{\sigma^2 + \|x_t\|^2_{H_t}} H_t x_t' \beta_{t|t-1} x_t - \frac{\|x_t\|^2_{H_t}}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \frac{y_t}{\sigma^2} x_t. \tag{B.80}$$

Multiplying the second and third term on the right-hand side with $\frac{\sigma^2 + \|x_t\|^2_{H_t}}{\sigma^2 + \|x_t\|^2_{H_t}}$ and $\frac{\sigma^2}{\sigma^2}$, respec-

tively, allows us to combine the second through fourth terms as follows

$$\beta_{t|t} = \beta_{t|t-1} + \frac{\sigma^2}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \frac{y_t - x_t' \beta_{t|t-1}}{\sigma^2} x_t, \tag{B.81}$$

where using the definition of the explicit gradient $\nabla(y_t|\beta_{t|t-1}, x_t)$ gives the final result

$$\beta_{t|t} = \beta_{t|t-1} + \frac{\sigma^2}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \nabla(y_t|\beta_{t|t-1}, x_t). \tag{B.82}$$