TI 2014-046/III
Tinbergen Institute Discussion Paper

Information Theoretic Optimality of Observation Driven Time Series Models

Francisco Blasques[a]
Siem Jan Koopman[a,b]
André Lucas[a]

[a] *Faculty of Economics and Business Administration, VU University Amsterdam, Tinbergen Institute, the Netherlands;*
[b] *CREATES, Aarhus University, Denmark.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Information Theoretic Optimality of Observation Driven Time Series Models

Francisco Blasques[a], Siem Jan Koopman[a,b], and André Lucas[a]

(a) *VU University Amsterdam and Tinbergen Institute*

(b) *CREATES, Aarhus University*

April 10, 2014

Abstract

We investigate the information theoretic optimality properties of the score function of the predictive likelihood as a device to update parameters in observation driven time-varying parameter models. The results provide a new theoretical justification for the class of generalized autoregressive score models, which covers the GARCH model as a special case. Our main contribution is to show that only parameter updates based on the score always reduce the local Kullback-Leibler divergence between the true conditional density and the model implied conditional density. This result holds irrespective of the severity of model misspecification. We also show that the use of the score leads to a considerably smaller global Kullback-Leibler divergence in empirically relevant settings. We illustrate the theory with an application to time-varying volatility models. We show that the reduction in Kullback-Leibler divergence across a range of different settings can be substantial in comparison to updates based on for example squared lagged observations.

1 Introduction

We provide the information theoretic optimality properties of a class of observation driven time-varying parameter models. The main distinguishing feature

of our model class is the use of the score function of the conditional (or the predictive) observation density to drive the changes of the parameters over time; see Creal, Koopman, and Lucas (2008, 2013) and Harvey (2013). Score driven models have been successfully applied in various empirical studies; see Appendix A for a short literature review. Furthermore, the score driven class of models encompasses many familiar observation driven time-varying parameter models such as the generalized autoregressive conditional heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986), the autoregressive conditional duration model of Engle and Russell (1998), the multiplicative error model of Engle (2002), and also the observation driven Poisson count model of Davis, Dunsmuir, and Streett (2003).

Despite the widespread use of these models, no firm theoretical foundation has currently been provided for the use of the score as a driving mechanism of the dynamics in observation driven time-varying parameter models. This paper aims to provide such a formal justification using an information theoretic perspective based on the Kullback-Leibler (KL) divergence. The KL divergence was first introduced by Kullback and Leibler (1951) as a measure of divergence between probability distributions, a work pioneered by Boltzmann in the 1870's with his concept of entropy in physics and thermodynamics. This work was extended by Shannon (1948). The importance of the KL divergence became increasingly clear when researchers started to uncover its key role in Fisher's information and sufficient statistics (Kullback, 1959), the maximum likelihood principle (Akaike, 1973), Laplace's principle of insufficient reason (Jaynes, 2003), the minimum discrimination principle or the principle of maximum entropy (Jaynes, 1957). Kapur and Kesavan (1992) and Cover and Thomas (1991) provide textbook treatments of applications in applied sciences. Maasoumi (1986) and Ullah (1996, 2002) review econometric applications. Given its fundamental and central place in information theory, machine learning, econometrics, statistical mechanics and many other fields, the KL divergence emerges as a natural starting point to develop a theoretical motivation for the use of the score function in observation

driven time series models.

To illustrate our objective, consider a (possibly multivariate) stochastic process $\{y_t\}_{t\in\mathbb{Z}}$ characterized by a true conditional density $y_t \sim p(y_t|f_t)$, where the unobserved time-varying parameter f_t is assumed to represent all dynamic features of y_t. For example, the time-varying variable f_t can represent a volatility, intensity, correlation, or copula dependence parameter. We postulate a statistical model that is characterized by a possibly misspecified conditional density $\tilde{p}(\cdot\,|\tilde{f}_t;\boldsymbol{\theta})$ to approximate the true conditional density $p(\cdot\,|f_t)$, where $\tilde{f}_t$ is a filtered estimate of the true f_t, and $\boldsymbol{\theta}\in\Theta$ is a static, unknown parameter vector. For example, $\tilde{f}_t$ may be the filtered volatility estimate from a GARCH model with a Student's t conditional density $\tilde{p}(\cdot\,|\tilde{f}_t;\boldsymbol{\theta})$, while f_t is the log volatility from a stochastic volatility (SV) model for a normal conditional density $p(\cdot\,|f_t)$; see, for example, Shephard (2005) for formulations of an SV model. A key feature of the predictive score modeling framework is that the dynamics of $\tilde{f}_t$ are driven by scaled versions of the score $\partial \ln \tilde{p}(y_t|\tilde{f}_t;\boldsymbol{\theta})/\partial\tilde{f}_t$, i.e., the derivative of the postulated log conditional observation density. For example, if $\tilde{p}(\cdot\,|\tilde{f}_t;\boldsymbol{\theta})$ is the normal density with mean zero and variance $\tilde{f}_t$ and the score is scaled by the inverse conditional Fisher information, we recover the familiar GARCH model of Engle (1982) and Bollerslev (1986). If, however, $\tilde{p}(\cdot\,|\tilde{f}_t;\boldsymbol{\theta})$ is a Student's t density, we do not recover the t-GARCH model of Bollerslev (1986), but an observation driven model for a time-varying variance with more robust properties than the t-GARCH model; see Creal et al. (2013) and Harvey (2013) for further details.

In this paper we refer to our class of predictive score driven time-varying parameter models as the generalized autoregressive score (GAS) model. We show that the parameter update in the GAS model is successful in reducing the Kullback-Leibler divergence between the true conditional density $p(y_t|f_t)$ and the model implied conditional density $\tilde{p}(y_t|\tilde{f}_t;\boldsymbol{\theta})$. In particular, the theoretical results in this paper reveal that the score is successful in ensuring that the GAS update reduces the KL divergence in expectation and at every step. Furthermore, we show that only score-driven updates can have this property. Since the score

makes use of 'local information' about the likelihood, the theory focuses naturally on 'local updates'. However, we show numerically that the optimality properties of the GAS hold more generally, even when large updates are present. All findings apply to correctly specified models as well as misspecified models.

Other arguments for the use of the score function in dynamic models have been proposed earlier in the filtering literature. For example, Masreliez (1975), Durbin and Koopman (1997) and Müller and Petalas (2010) motivate the use of the score for filtering based on a Laplace approximation of the postulated parametric observation density in the state space framework. The scores are then used to build an approximating linear Gaussian state space model for the efficient estimation of dynamic parameters via the Kalman filter, possibly combined with importance sampling techniques; see Durbin and Koopman (2012) for a recent textbook treatment. Another argument is found in Nelson and Foster (1994). Using fill-in asymptotics and a near-diffusion framework, they show that filters based on the score of the observation density are optimal in a mean-squared-error sense when estimating the true, unobserved f_t. Compared to this earlier literature, our paper takes the perspective of optimality from an information theoretic point of view. Information theoretic arguments provide one of the common benchmarks for most econometricians and statisticians by which to assess the statistical adequacy of alternative procedures.

The remainder of this paper is organized in the following way. In Section 2, we introduce the main concepts and definitions used to evaluate the information theoretic properties of the GAS updating scheme. In Section 3, we formulate the key propositions that establish the local and global optimality of GAS updates. In Section 4, we provide some illustrative examples for time-varying volatility models to highlight the main aspects of our theoretical results. We conclude in Section 5. Appendix A provides a short empirical literature review on predictive score driven time-varying parameter models while Appendix B gathers the proofs.

2 Notation and optimality concepts

We consider a stochastic process $\{y_t\}_{t\in\mathbb{Z}}$ with elements taking values in $\mathcal{Y} \subseteq \mathbb{R}$. The process is characterized by a true conditional density $y_t \sim p(y_t|f_t)$, where $f_t \in \mathcal{F} \subset \mathbb{R}$ is unobserved. We observe a given sequence of T observations, $y^T := \{y_t\}_{t=1}^T$, and consider the statistical model given by

$$y_t \sim \tilde{p}(y_t|\tilde{f}_t;\boldsymbol{\theta}), \tag{1}$$

$$\tilde{f}_{t+1} = \phi\Big(\tilde{f}_t, y_t\,;\,\boldsymbol{\theta}\Big), \tag{2}$$

where $\tilde{p}(\cdot|\tilde{f}_t;\boldsymbol{\theta})$ is a parametric conditional density, $\tilde{f}_t \in \tilde{\mathcal{F}} \subset \mathbb{R}$ is a filtered value of the true f_t, $\boldsymbol{\theta} \in \Theta$ is a vector of static, unknown parameters, and ϕ is an updating function linking the new $\tilde{f}_{t+1}$ to the current observation y_t and the current filtered time-varying parameter $\tilde{f}_t$. The model in equations (1)–(2) implies that we adopt an observation driven approach in our statistical analysis; see Cox (1981) for a detailed description. In particular, we assume that $\tilde{f}_t$ is a measurable function of y^{t-1} and $\boldsymbol{\theta}$, i.e., $\tilde{f}_t = \tilde{f}_t(y^{t-1}, \boldsymbol{\theta}, \bar{f}_1)$ for some initial value $\bar{f}_1$. For example, in the familiar GARCH setting, ϕ takes the form $\phi(\tilde{f}_t, y_t;\boldsymbol{\theta}) = \omega^* + \alpha^* y_t^2 + \beta^* \tilde{f}_t$ for a fixed unknown parameter vector $\boldsymbol{\theta} = (\omega^*, \alpha^*, \beta^*)'$.

The approximation of $p(\cdot\,|f_t)$ by $\tilde{p}(\cdot\,|\tilde{f}_t)$ in (1) is two-fold, as it relates to the shape of the density function itself as well as to the filtered time-varying parameter $\tilde{f}_t$. Both features may possibly depend on the parameter vector $\boldsymbol{\theta}$, which needs to be estimated. The estimation of $\boldsymbol{\theta}$ is therefore another key part of our analysis.

The generalized autoregressive score (GAS) framework also falls in the class defined by equations (1)–(2). The defining property of GAS models is the use of the score in updating $\tilde{f}_t$ to $\tilde{f}_{t+1}$. For example, for a first order autoregressive

scheme the GAS model is given by

$$\tilde{f}_{t+1} = \phi\Big(\tilde{f}_t, y_t\,;\,\boldsymbol{\theta}\Big) = \omega + \alpha \cdot S_t \cdot \tilde{\nabla}_t + \beta \tilde{f}_t, \tag{3}$$

$$\tilde{\nabla}_t := \tilde{\nabla}\Big(\tilde{f}_t, y_t\,;\,\boldsymbol{\theta}\Big) = \partial \log \tilde{p}\Big(y_t\Big|f;\boldsymbol{\theta}\Big)/\partial f\,\Big|_{f=\tilde{f}_t}, \tag{4}$$

where $S_t := S(\tilde{f}_t;\boldsymbol{\theta})$ is a positive scaling function that possibly depends on the filtered time-varying parameter $\tilde{f}_t$ and the static parameter $\boldsymbol{\theta}$. For example, Creal et al. (2013) propose to scale by powers of the conditional Fisher information to account for the curvature of the score function. We assume throughout this paper that $\tilde{p}(\cdot|\tilde{f})$ is continuously differentiable in $\tilde{f} \in \tilde{\mathcal{F}}$, and that the score function $\tilde{\nabla}_t$ in (4) is continuously differentiable in both $\tilde{f}_t$ and y_t. We further assume that $\tilde{f}_{t+1}$ is a continuous random variable that has a density, such that the model does not degenerate; see, for example, Blasques et al. (2012, 2014) for more precise conditions on the stationarity and ergodicity of $\tilde{f}_t$.

The key objective of our study is to characterize functions ϕ that possess optimality properties from an information theoretic point of view. To achieve this objective, we first define a number of relevant optimality properties that we analyze in more detail below.

Given a true unobserved sequence $\{f_t\}$ and an approximate filtered sequence $\{\tilde{f}_t\} = \{\tilde{f}_t(y^{t-1}, \boldsymbol{\theta}, \bar{f}_1)\}$, the optimal information-theoretic update of $\tilde{f}_t$ to a new value $\tilde{f}_{t+1}$ minimizes the Kullback-Leibler (KL) divergence between the true conditional density $p_t := p(\cdot\,|f_t)$ and the postulated density $\tilde{p}_{t+1} := \tilde{p}(\cdot\,|\tilde{f}_{t+1};\boldsymbol{\theta})$. The KL distance is then defined as

$$\mathcal{D}_{\mathrm{KL}}\Big(p_t\,,\,\tilde{p}_{t+1}\Big) = \int_Y p(y|f_t) \ln \frac{p(y|f_t)}{\tilde{p}(y|\tilde{f}_{t+1};\boldsymbol{\theta})}\,\mathrm{d}y, \tag{5}$$

where $Y \subseteq \mathbb{R}$ is the subset of the real line over which the divergence is evaluated. Selecting Y to be a small neighborhood of a point yields a local KL-divergence as found, for example, in Hjort and Jones (1996) and Ullah (2002). In these contributions, local ML estimation is related to the minimization of the local KL divergence or the maximization of the local Shannon entropy. If $Y = \mathbb{R}$, then

$\mathcal{D}_{\mathrm{KL}}$ corresponds to global KL divergence of Kullback and Leibler (1951).

The $\mathcal{D}_{\mathrm{KL}}$ optimality concept follows from the principle of Minimum Discrimination Information (MDI) proposed by Kullback (1959). When a new observation y_t becomes available, we choose $\tilde{f}_{t+1}$ such that the updated density $\tilde{p}_{t+1}$ is as hard to discriminate over the domain Y from the true density p_t as possible. In other words, the new observation should produce as small an information gain $\mathcal{D}_{\mathrm{KL}}(p_t, \tilde{p}_{t+1})$ in (5) as possible. The KL distance in (5) can also be expressed as

$$\mathcal{D}_{\mathrm{KL}}\Big(p_t, \tilde{p}_{t+1}\Big) = \int_Y p(y|f_t)\ \ln p(y|f_t)\,\mathrm{d}y\ -\ \int_Y p(y|f_t)\ \ln \tilde{p}(y|\tilde{f}_{t+1};\boldsymbol{\theta})\,\mathrm{d}y, \qquad (6)$$

where the first term on the right-hand side of (6) corresponds to the information entropy of the true density p_t, and where the second term corresponds to the cross entropy between the true density p_t and the approximate density $\tilde{p}_{t+1}$. Minimizing the KL divergence thus corresponds to maximizing the cross entropy.

The most natural optimality concept that arises from the KL divergence is one that defines a parameter update as optimal if it minimizes the KL divergence between the true conditional density p_t and the modeled conditional density $\tilde{p}_{t+1}$,

$$\tilde{f}_{t+1} \in \arg\min_{f\in\mathcal{F}} \mathcal{D}_{\mathrm{KL}}\Big(p_t\ ,\ \tilde{p}(\,\cdot\,|f;\boldsymbol{\theta})\Big). \qquad (7)$$

However, we can only apply this concept of optimality when the conditional density p is known and the sequence $\{f_t\}$ is observed. In empirical work, this concept is infeasible in most situations of practical interest. Therefore, we turn to alternative notions of optimality that (i) can be applied in empirical settings where the data generating process (DGP) is unknown, (ii) can provide an important characterization of any observation driven parameter updating scheme, and (iii) can build on the solid foundations of KL optimality as discussed above.

An important notion of optimality that satisfies the above criteria focuses on the *improvement* that the updating step produces in terms of the KL divergence from the true conditional density. In particular, given a starting point for the filtered parameter $\tilde{f}_t$ and a conditional density $\tilde{p}(\cdot|\tilde{f}_t)$, we analyze the conditions

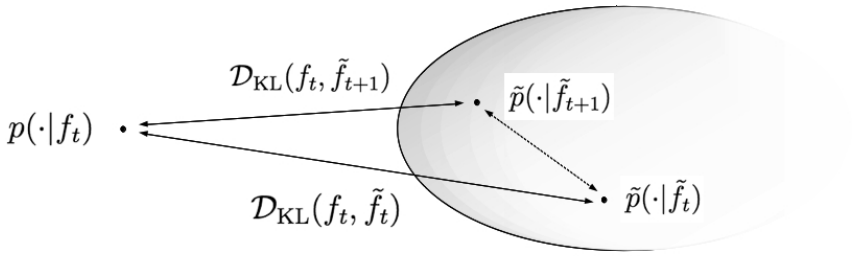

Figure 1: A graphical representation of first-order optimality under model misspecification. The shaded area denotes the set of conditional densities generated by the model.

under which the new observation y_t, drawn from $p(\cdot|f_t)$, produces an update from $\tilde{f}_t$ to $\tilde{f}_{t+1}$ such that the new conditional density $\tilde{p}(\cdot|\tilde{f}_{t+1})$ provides a better approximation to $p(\cdot|f_t)$ than $\tilde{p}(\cdot|\tilde{f}_t)$. Hence we focus on the change in KL divergence resulting from updating the time-varying parameter from a point $\tilde{f}_t \in \tilde{\mathcal{F}}$ to another point $\tilde{f}_{t+1} \in \tilde{\mathcal{F}}$. We call this change the realized KL variation.

Definition 1. (RKL Optimality) *The realized KL (RKL) variation of a parameter update from $\tilde{f}_t \in \tilde{\mathcal{F}}$ to $\tilde{f}_{t+1} \in \tilde{\mathcal{F}}$ is given by*

$$\begin{aligned}\Delta_{t|t} &= \mathcal{D}_{\mathrm{KL}}\Big(p_t, \tilde{p}_{t+1}\Big) - \mathcal{D}_{\mathrm{KL}}\Big(p_t, \tilde{p}_t\Big) \\ &= \int_Y p(y|f_t) \ \Big(\ln \tilde{p}(y|\tilde{f}_t; \boldsymbol{\theta}) - \ln \tilde{p}(y|\tilde{f}_{t+1})\Big) \ \mathrm{d}y.\end{aligned}$$

For a given p_t, a parameter update is RKL optimal if and only if $\Delta_{t|t} < 0$.

The KL divergences in the definition of $\Delta_{t|t}$ are both taken with respect to the same unknown density $p_t \equiv p(y_t|f_t)$ because the realized observation y_t, used in updating $\tilde{p}_t$ to $\tilde{p}_{t+1}$, is drawn from p_t. In other words, starting from $\tilde{p}_t$, the observation driven update can only use the draw from p_t to approximate p_t itself more accurately. Figure 1 illustrates, in a graphical way, a realized update that is KLV optimal despite model misspecification.

In a dynamic system that is subject to stochastic perturbations, it is not always possible to ensure that the realized step is optimal. The GAS update is no exception. For example, in a time-varying volatility model, it is conceivable that an outlier materializes in a period when the true volatility has gone down. From an inferential perspective, however, the outlying observation naturally leads to an

increase in our volatility estimate. This is inherently difficult to avoid. However, an optimal updating scheme, while being subject to stochastic perturbations, should have a tendency to move in the 'correct direction' on average in the sense of the KL divergence reducing in expectation. For this purpose we introduce the concept of conditionally expected KL optimality.

DEFINITION 2. (CKL Optimality) *The conditionally expected KL (CKL) variation of a parameter update from* $\tilde{f}_t \in \tilde{\mathcal{F}}$ *to* $\tilde{f}_{t+1} \in \tilde{\mathcal{F}}$ *is given by*

$$\Delta_{t|t-1} = \int_F q(\tilde{f}_{t+1}|\tilde{f}_t, f_t; \boldsymbol{\theta}) \left[\int_Y p(y|f_t) \ln \frac{\tilde{p}(y|\tilde{f}_t; \boldsymbol{\theta})}{\tilde{p}(y|\tilde{f}_{t+1}; \boldsymbol{\theta})} \mathrm{d}y \right] d\tilde{f}_{t+1},$$

where $q(\tilde{f}_{t+1}|\tilde{f}_t, f_t; \boldsymbol{\theta})$ *denotes the density of* $\tilde{f}_{t+1}$ *conditional on both* $\tilde{f}_t$ *and* f_t*. For a given* p_t*, an update is CKL optimal if and only if* $\Delta_{t|t-1} < 0$*.*

Conditional on information available up to time $t-1$, the filtered parameter $\tilde{f}_t$ is known and fixed because $\tilde{f}_t \equiv \tilde{f}_t(y^{t-1}, \boldsymbol{\theta}, \bar{f}_1)$. The parameter $\tilde{f}_{t+1}$ by contrast is unknown and random because $\tilde{f}_{t+1} \equiv \tilde{f}_{t+1}(y^t, \boldsymbol{\theta}, \bar{f}_1)$ depends on the unrealized y_t, which is random even for a given f_t. The true unknown measure of y_t plays a role in the definition of CKL optimality through the conditional density $p(y|f_t)$ in the inner integral, as well as through the conditional density $q(\tilde{f}_{t+1}|\tilde{f}_t, f_t; \boldsymbol{\theta})$ in the outer integral. Despite this dependence on the unknown true conditional density and f_t, we can still establish theoretical properties based on RKL and CKL optimality.

3 Optimality of GAS updates

3.1 Local optimality of GAS updates

We first establish a number of analytical results for local optimality that hold for every true density p_t. Local results focus on the 'direction' of the updating step. A locally optimal update must be in a correct direction, i.e., in a direction that reduces the KL divergence.

Our first results show that the updates from the GAS models are locally RKL and CKL optimal for every p_t. In particular, the optimality proofs show that $\Delta_{t-1|t-1} < 0$ and $\Delta_{t|t-1} < 0$ for every p_t on a neighborhood $\tilde{f}_t$ and y_t, respectively. In other words, we show that $\exists \delta_f > 0 \wedge \delta_y > 0$ such that $\Delta_{t-1|t-1} < 0$ and $\Delta_{t|t-1} < 0$ holds for every p_t on the sets of the form

$$\begin{aligned} F &= F_{\delta_f}(\tilde{f}_t) := \{\tilde{f} \in \tilde{\mathcal{F}} : |\tilde{f} - \tilde{f}_t| < \delta_f\}, \\ Y &= Y_{\delta_y}(y_t) := \{y \in \mathcal{Y} : |y - y_t| < \delta_y\}. \end{aligned} \tag{8}$$

The set Y may be chosen more effectively for particular models. For example, in symmetric volatility models, we may want to use the distance between $|y|$ and $|y_t|$ so that our local optimality results hold for even a larger set of Y. We abstract from such alternative choices of distances to avoid notational burden.

We require some regularity conditions on the support of the true conditional density p_t and the score of the postulated model $\tilde{\nabla}_t$. These conditions can be relaxed easily at the cost of more complex notation and proofs; we mainly require the existence of regions of positive measures for the observed data y_t over which the model score $\tilde{\nabla}_t$ is well defined and non-zero.

Assumption 1. *$p(y|f) > 0 \ \forall \ (y, f) \in \mathbb{R} \times \mathcal{F}$ and $\tilde{\nabla}(\tilde{f}, y; \boldsymbol{\theta}) \neq 0$ for every $(\tilde{f}, \boldsymbol{\theta}) \in \tilde{\mathcal{F}} \times \Theta$ and almost every $y \in \mathbb{R}$.*

Assumption 2. *$\alpha > 0$ and $S(\tilde{f}; \boldsymbol{\theta}) > 0 \ \forall \ (\tilde{f}, \boldsymbol{\theta}) \in \tilde{\mathcal{F}} \times \Theta$.*

The first condition in Assumption 1 excludes those values of the time-varying parameter that result in a distribution for y_t that is degenerate. For example, in the volatility models of Section 4, we exclude the possibility of the variance becoming zero. The second condition in Assumption 1 rules out the possibility of the time-varying parameter being non-identified at certain update steps. The Assumption 2 imposes two trivial conditions that ensure that the GAS update does not 'ignore' or 'distort' the information contained in the score.

Definition 3. (Newton-GAS update) *The Newton-GAS update is defined as* (3) *with $\omega = 0$ and $\beta = 1$, from which we obtain $\tilde{f}_{t+1} = \tilde{f}_t + \alpha \tilde{\nabla}_t$.*

The Newton-GAS update is our building block in the subsequent analysis. It resembles the update in the numerical Newton algorithm using steepest ascent steps.

PROPOSITION 1. *Let Assumptions 1 and 2 hold. Then, every Newton-GAS update is locally RKL optimal and CKL optimal for any true density* p_t.

Proposition 1 above shows that the parameter restrictions $\alpha > 0 \wedge (\omega, \beta) = (0, 1)$ ensure that $\Delta_{t|t-1} < 0$ and $\Delta_{t|t} < 0$ hold for every p_t and every $(\tilde{f}_{t+1}, y)$ in a neighborhood of $(\tilde{f}_t, y_t)$. This result can be achieved because, locally, the sign of the score indicates correctly whether the time-varying parameter should be updated upwards or downwards. Proposition 2 below shows that the properties derived above are only available for 'score-equivalent' updates. It establishes that we require a fundamental sign condition for an update to achieve the results of Proposition 1: the update must give exactly the same local information as the score in the sense that their signs must be the same.

DEFINITION 4. (Score-Equivalent Update) *The observation driven parameter update (2) is 'score-equivalent' if and only if* $\text{sign}(\Delta\phi(f, y; \boldsymbol{\theta})) = \text{sign}(\tilde{\nabla}(f, y; \boldsymbol{\theta}))$ *for almost every* $(y, f) \in \mathcal{Y} \times \mathcal{F}$ *and every* $\boldsymbol{\theta}$.

PROPOSITION 2. *Let Assumptions 1 and 2 hold. For any given true density* p_t, *a parameter update is locally RKL optimal and CKL optimal if and only if the parameter update is score-equivalent.*

Proposition 2 underlines the importance of the score for KL related optimality concepts. The property stated in the proposition, however, holds for a larger class of GAS updates than just those driven purely by the score. In particular, we allow $(\omega, \beta) \neq (0, 1)$ as long as the 'force away' from the optimal direction at $\tilde{f}_t$, which is determined by the autoregressive component $\omega + (\beta - 1)\tilde{f}_t$, is weaker than the 'force towards' the optimal direction, which is determined by the score component $\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})$. Indeed, the optimality regions defined in our next proposition reveal that, for any given $\tilde{f}_t = \tilde{f} \in \tilde{\mathcal{F}}$, both RKL and CKL

optimality hold as long as α is larger than a multiple of $\omega + (\beta - 1)\tilde{f}$. This is summarized in Proposition 3.

PROPOSITION 3. *Let Assumptions 1 and 2 hold. Then, the GAS update is locally RKL optimal for every p_t if*

$$\alpha > \frac{|\omega + (\beta - 1)\tilde{f}_t|}{S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|}. \tag{9}$$

The GAS update is locally CKL optimal for every p_t if

$$\alpha > \frac{\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|}{S(\tilde{f}_t; \boldsymbol{\theta})\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|^2}|\omega + (\beta - 1)\tilde{f}_t|, \tag{10}$$

with $\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})| = \int_{Y_f} p(y_t|f_t)\,|\tilde{\nabla}(\tilde{f}_t, y; \boldsymbol{\theta})|\mathrm{d}y$ *and* $Y_f := \{y \in \mathbb{R} : |\phi(\tilde{f}_t, y; \boldsymbol{\theta}) - \tilde{f}_t| < \delta_f\}$.

The parameter restrictions in Proposition 3 show precisely how large α must be for the score to have a 'greater influence' on the parameter update than the autoregressive part. Equation (9) can be rewritten as

$$\alpha S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})| > |\omega + (\beta - 1)\tilde{f}_t|,$$

and it shows directly that two forces play a role in local optimality. For any given value of α, the larger the absolute value of the scaled score $S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|$, the more likely we have a realized step that is locally optimal. Similarly, the closer ω is to zero or the closer β is to one, which corresponds to the Newton-GAS update, the easier it is to obtain local optimality. For RKL optimality, the intuition follows when taking an arbitrarily small value for δ_f in (8): the change $\tilde{f}_{t+1} - \tilde{f}_t$ is then mainly driven by the score part, and by concentrating on a small enough neighborhood $F = F_{\delta_f}(\tilde{f}_t)$, the expression for RKL-variation becomes (negative) quadratic.

EXAMPLE 1. For the GAS volatility model with a normal distribution, we obtain a model equivalent to the standard GARCH model. We can reduce equation (9)

to $\alpha\,|y_t^2 - \tilde{f}_t| > |\omega + (\beta - 1)\tilde{f}_t|$, such that the update is convincingly more RKL optimal if the observed y_t^2 deviates considerably from the filtered volatility $\tilde{f}_t$.

The parameter restrictions for CKL optimality have a similar interpretation. In particular, although the local conditional expectations in the parameter restriction (10) depend on the unknown conditional density p_t, the condition is valid for every conditional density p_t because by Assumption 2 the denominator cannot be zero. Hence the restriction in (10) imposes the required condition that the optimality result holds for every p_t.

3.2 Non-local optimality of GAS updates

In this section we extend the local results to a non-local setting. These results are not only concerned with the direction of the updating step, but also, with the size of the updating step. In other words, we characterize the step size for which we can ensure optimality.

It is convenient to introduce notation for the local supremum of the partial derivatives of the score $\tilde{\nabla}_t$,

$$\begin{aligned}
\xi_{\delta_f,\delta_y}\left(\tilde{f}_t, y_t\right) &:= \sup_{(f,y)\in F_{\delta_f}(\tilde{f}_t)\times Y_{\delta_y}(y_t)} \left|\frac{\partial\tilde{\nabla}(f,y;\boldsymbol{\theta})}{\partial f}\right|, \\
\zeta_{\delta_f,\delta_y}\left(\tilde{f}_t, y_t\right) &:= \sup_{(f,y)\in F_{\delta_f}(\tilde{f}_t)\times Y_{\delta_y}(y_t)} \left|\frac{\partial\tilde{\nabla}(f,y;\boldsymbol{\theta})}{\partial y}\right|.
\end{aligned}$$

In Proposition 1 we established the local optimality of the Newton-GAS update. In Proposition 4 we show that the Newton-GAS is optimal on a larger set as long as this set satisfies some size restrictions.

Proposition 4. *Let Assumptions 1 and 2 hold. Then, the Newton-GAS update is RKL optimal on sets $F_{\delta_f}(\tilde{f}_t)$ and $Y_{\delta_y}(y_t)$ that satisfy*

$$\eta_{\delta_f,\delta_y}(\tilde{f}_t, y_t) < |\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|, \tag{11}$$

and CKL optimal on sets that satisfy

$$\eta_{\delta_f,\delta_y}(\tilde{f}_t, y_t) < \frac{\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|^2}{\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|}, \tag{12}$$

where $\eta_{\delta_f,\delta_y}(\tilde{f}_t, y_t) := \xi_{\delta_f,\delta_y}(\tilde{f}_t, y_t) \times \delta_f + \zeta_{\delta_f,\delta_y}(\tilde{f}_t, y_t) \times \delta_y$.

Example 1 (continued). Consider the model $y_t = \tilde{f}_t \epsilon_t$ where ϵ_t comes from a $\chi^2(1)$ distribution. This model is again equivalent to the GARCH model with normal disturbances, where we have taken the squares of both sides of the equation $\tilde{y}_t = \tilde{f}_t^{1/2}\tilde{\epsilon}_t$, with $y_t = \tilde{y}_t^2$ and $\tilde{\epsilon}_t \sim \mathrm{N}(0,1)$. The GAS model with inverse Fisher information scaling of the score is again equivalent to the GARCH(1,1). We have $\tilde{f}_{t+1} = \omega + \beta\tilde{f}_t + \alpha(y_t - \tilde{f}_t) = \omega + \beta\tilde{f}_t + \alpha(\tilde{y}_t^2 - \tilde{f}_t)$, and $\xi_{\delta_f,\delta_y}(\tilde{f}_t, y_t; \boldsymbol{\theta}) = 1$ and $\zeta_{\delta_f,\delta_y}(\tilde{f}_t, y_t; \boldsymbol{\theta}) = 1$. Restriction (11) then takes the form $\delta_f + \delta_y < |y_t - \tilde{f}_t|$. This region is thus larger if the observed y_t is at odds with the last filtered variance $\tilde{f}_t$.

Example 2. If we consider the Student's t GAS volatility model of Creal et al. (2013) and Harvey (2013), the functions $\xi_{\delta_f,\delta_y}(\tilde{f}_t, y_t)$ and $\zeta_{\delta_f,\delta_y}(\tilde{f}_t, y_t)$ are not uniformly bounded. Restriction (11) takes the form

$$\xi_{\delta_f,\delta_y}\left(\tilde{f}_t, y_t\right) \times \delta_f + \zeta_{\delta_f,\delta_y}\left(\tilde{f}_t, y_t\right) \times \delta_y \;<\; \left|1 + 3\lambda^{-1}\right| \left| \frac{(1+\lambda^{-1})\, y_t^2}{1 + \lambda^{-1} y_t^2/\tilde{f}_t} - \tilde{f}_t \right|, \tag{13}$$

where λ denotes the degrees of freedom parameter of the Student's t distribution, and both $\xi_{\delta_f,\delta_y}(\tilde{f}_t, y_t)$ and $\zeta_{\delta_f,\delta_y}(\tilde{f}_t, y_t)$ are increasing in $\tilde{f}_t$ and decreasing in y_t^2 and λ.

The optimality regions described above are obtained analytically from sufficient conditions. In Section 4 we provide a numerical description of the true optimality regions by dealing with a case where the DGP is known.

Finally, for general GAS updates with $(\omega, \beta) \neq (0, 1)$, we obtain the following optimality result.

PROPOSITION 5. *Let Assumptions 1 and 2 hold. Then, the GAS update is RKL optimal on sets $F_{\delta_f}(\tilde{f}_t)$ and $Y_{\delta_y}(y_t)$ that satisfy*

$$\begin{aligned}\alpha S(\tilde{f}_t;\boldsymbol{\theta})\,\tilde{\nabla}(\tilde{f}_t,y_t;\boldsymbol{\theta})^2 > &\Big(\eta_{\delta_f,\delta_y}\big(\tilde{f}_t,y_t\big)+\Big|\tilde{\nabla}(\tilde{f}_t,y_t;\boldsymbol{\theta})\Big|\Big)\,|\omega+(\beta-1)\tilde{f}_t|+\\ &\alpha S(\tilde{f}_t;\boldsymbol{\theta})\,\eta_{\delta_f,\delta_y}(\tilde{f}_t,y_t)\Big)\,\Big|\tilde{\nabla}(\tilde{f}_t,y_t;\boldsymbol{\theta})\Big|,\end{aligned} \tag{14}$$

and CKL optimal on sets that satisfy

$$\begin{aligned}\alpha S(\tilde{f}_t;\boldsymbol{\theta})\,\mathbb{E}^{t-1}_{Y_f}\tilde{\nabla}(\tilde{f}_t,y_t;\boldsymbol{\theta})^2 > &\Big(\eta_{\delta_f,\delta_y}\big(\tilde{f}_t,y_t\big)+\mathbb{E}^{t-1}_{Y_f}\Big|\tilde{\nabla}(\tilde{f}_t,y_t;\boldsymbol{\theta})\Big|\Big)\,|\omega+(\beta-1)\tilde{f}_t|+\\ &\alpha S(\tilde{f}_t;\boldsymbol{\theta})\,\eta_{\delta_f,\delta_y}(\tilde{f}_t,y_t)\,\mathbb{E}^{t-1}_{Y_f}\Big|\tilde{\nabla}(\tilde{f}_t,y_t;\boldsymbol{\theta})\Big|.\end{aligned}$$

Compared to the local results of Proposition 3 that assumed arbitrarily small updating steps, the results of Proposition 5 allow the size of updating steps to be larger, as long as the appropriate bounds are satisfied. Note that the conditions in Proposition 5 collapse to those in Proposition 3 if $\omega = 0$ and $\beta = 1$.

To conclude, we re-emphasize that the optimality regions derived in this section are valid for any postulated conditional density $\tilde{p}$, regardless of the true unknown DGP p. At the same time, however, all model density choices $\tilde{p}$ are *not* the same. In order to minimize the KL divergence it is always preferable to work with a model density that approximates the DGP best. Hence, a good choice for the conditional density remains crucial. The current results only state that for a given choice of model density, the dynamics of $\tilde{f}_t$ as driven by the score of $\tilde{p}$ itself possess a number of information theoretic optimality properties. Observation driven updates based on other mechanisms do not necessarily have similar properties.

To make the last point more concrete, we consider three models in the next section: a normal distribution $\tilde{p}$ and score based updates (i.e., the normal GARCH or GAS model), a Student's t density $\tilde{p}$ with GARCH updates (t-GARCH), and a Student's t density with score updates (t-GAS). Both the GARCH and t-GAS model are optimal in the sense defined in this section. For a fat-tailed DGP, however, the non-optimal t-GARCH performs better in general

than the 'optimal' normal GARCH model. The reason is that the Student's t density $\tilde{p}$ of the t-GARCH fits the fat-tailed DGP density p much better than does the normal distribution of the normal GARCH model. However, given the choice for the Student's t model density $\tilde{p}$, the score based steps of the optimal t-GAS model perform better in terms of KL reductions than the steps of the t-GARCH model.

4 Application: volatility modeling

Due to their analytical nature, the results derived in Section 3 made use of sufficient conditions that bound the size of the regions F and Y over which we could ensure the optimality of the update for the time-varying parameter estimate. Here we take a numerical perspective and illustrate our KL optimality results for the stochastic volatility model. This allows us to numerically analyze the shape of the true optimality regions. Furthermore, it allows us to calculate KL divergences for $F = Y = \mathbb{R}$, such that effectively we can focus on global rather than local optimality in our numerical analysis. The illustration provides insights into whether the gains by using the score are only local, or also applicable for much larger regions of the sample and parameter space.

4.1 Observation driven approximations of the SV model

Time-varying volatility and fat tails are salient features in many financial time series; see, amongst others, Bollerslev (1986). Therefore, in our simulations we use a fat-tailed stochastic volatility (SV) model as a DGP, see e.g. Shephard (2005) and Durbin and Koopman (2012) for formulations of such SV models. We represent the DGP as

$$\begin{aligned} y_t &= \sqrt{f_t} u_t, & u_t &\sim p_u(u_t; \lambda), \\ \log f_t &= \mu + \rho \log f_{t-1} + \epsilon_t, & \epsilon_t &\sim \text{NID}(0, \sigma_\epsilon^2), \end{aligned} \tag{15}$$

Model	$p_u(u_t, \lambda)$	Update Equation
GARCH	$N(0,1)$	$\tilde{f}_{t+1} = \omega + \alpha(y_t^2 - \tilde{f}_t) + \beta\tilde{f}_t$
t-GARCH	$t(\lambda)$	$\tilde{f}_{t+1} = \omega + \alpha(y_t^2 - \tilde{f}_t) + \beta\tilde{f}_t$
t-GAS	$t(\lambda)$	$\tilde{f}_{t+1} = \omega + \alpha(1+3\tilde{\lambda})(w_t y_t^2 - \tilde{f}_t) + \beta\tilde{f}_t$

Table 1: Update functions for GARCH, t-GARCH and t-GAS models where $w_t^{-1} = (1 - 2\lambda)(1 + \lambda y_t^2/[(1-2\lambda)\tilde{f}_t])/(1+\lambda)$; see Creal et al. (2013) and Harvey (2013) for more detailed discussions.

for $t = 1, \ldots, T$, where T is the number of observations, f_t represent the true time-varying volatility, u_t is the standardized innovation with its corresponding Student's t distribution $p_u(\,\cdot\,;\lambda)$ and degrees of freedom λ, and $\{\epsilon_t\}$ is an independent normally distributed sequence with mean zero and constant variance σ_ϵ^2. We do not consider the leverage effect in our numerical example in this section and assume that u_t and ϵ_t are serially and mutually independent. The parameters of the SV model are set to empirically relevant values, $\mu = 0$, $\rho = 0.98$, $\sigma_\epsilon = 0.065$, and we let λ range from 3 to 8.

We consider three approximating observation driven models for (15): the standard GARCH model, which coincides with the GAS volatility model based on the normal distribution; the GARCH model with conditional Student's t distribution of Bollerslev (1986); and the GAS volatility model with conditional Student's t distribution of Creal et al. (2013) and Harvey (2013) as discussed in Section 2. These models share a common observation equation $y_t = \sqrt{\tilde{f}_t}u_t$ but impose different distributions for the innovations and make use of different updating equations for filtering the time-varying parameter $\tilde{f}_t$. Table 1 below offers establishes the notation.

To determine the pseudo-true parameter vector $\boldsymbol{\theta}$ that provides the best possible approximation to the DGP (15) in terms of KL divergence, we simulate a time series for y_t of length $T = 35,000$ and estimate the parameters of each model and DGP by maximum likelihood (ML). The consistency and asymptotic normality of the ML estimator to a pseudo-true parameter under model misspecification is ensured by the results of Blasques et al. (2014). We have verified that

for this large sample size the estimates of the pseudo-true values in do not change substantially if we add more observations or if we use a different random seed for simulation. All three observation driven models take $\tilde{f}_t$ as the squared scale parameter of the Student's t distribution. As a result, all models are misspecified in two dimensions: (i) the statistical models are observation driven, whereas the DGP is parameter driven, and (ii) the DGP dynamics are in the log scale parameter rather than in the squared scale parameter.

4.2 Estimation results

Figure 2 presents the estimation results. The strongest differences appear for the fat-tailed distributions; these are the models with low values of λ.

Since both the GARCH and t-GARCH models are highly sensitive to outliers, the pseudo-true α parameter tends to decrease as the true λ decreases and the innovations become fatter tailed. In this way the pseudo-true α of both the GARCH and t-GARCH model attempts to compensate for the fact that outliers in y_t produce spikes in the filtered volatility $\tilde{f}_t$ as the update equation uses y_t^2 to drive the dynamics of $\tilde{f}_t$. This effect is more noticeable in the t-GARCH because the innovations are already fat tailed. In the case of the GARCH model, the values of the pseudo-true parameters reflect also the misspecification in the conditional density, which is normal instead of Student's t. The GARCH pseudo-true parameters are closer to those of the t-GARCH when the true λ is larger. However, as the true λ decreases and the true density becomes fatter tailed, the GARCH model can only overcome the misspecification of its implied conditional density by changing its pseudo-parameters such that the GARCH filter produces a higher volatility. For very low λ, this is achieved by having larger values for the pseudo-true ω and α.

In comparison to the GARCH model, the t-GARCH model has a well specified conditional density that allows for fat tails in the innovations. However, since the t-GARCH uses y_t^2 to drive the dynamics of $\tilde{f}_t$, the volatility update is still very sensitive to outliers. In particular, when $|y_t|$ is very large, it causes a large

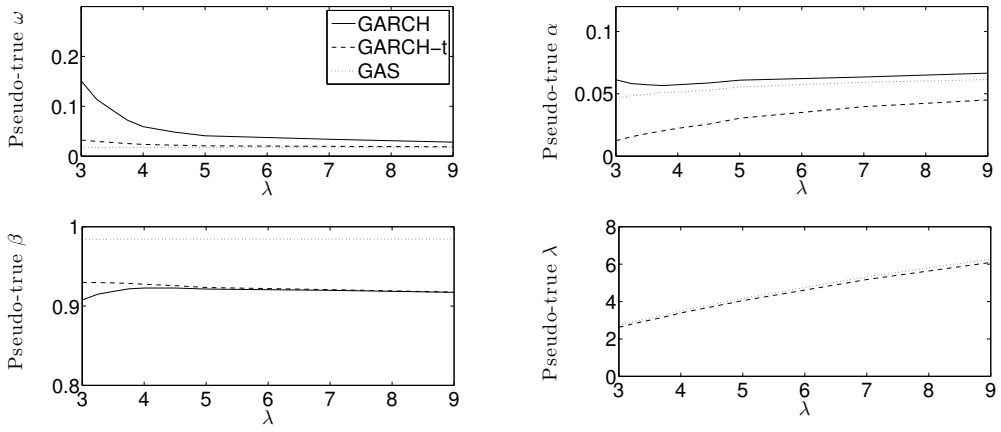

Figure 2: The pseudo-true parameters for GARCH, t-GARCH and t-GAS models using the stochastic volatility DGP (15) with $a = 0$, $b = 0.98$, $\sigma_\epsilon = 0.065$ and $\lambda \in [3, 8]$. The pseudo-true values are obtained by estimating each model separately for each true value of λ by maximum likelihood on a simulated time series of $T = 35,000$ observations.

spike in the filtered volatility estimate. To minimize this effect, the pseudo-true values for α are always decreasing in the true λ. However, a lower value of α in the t-GARCH model comes at the cost of the filtered volatility being lower on average and being unable to react to true increases of volatility. For very low values of λ, the pseudo-true value of ω then rises to compensate for this defect.

The GAS model does not suffer from these parameter adjustments because the impact of large y_t's, in absolute values, is downweighted in (13). As a result, its pseudo-true parameters reveal greater stability throughout the range of λ. In this sense, the GAS model is naturally more robust to outliers.

4.3 Kullback-Leibler divergence comparisons

In the left-hand panel of Figure 3 we present the relative difference in KL-divergence of the t-GAS model to that of the GARCH and t-GARCH models. Suppose that KL(G) denotes the KL divergence between the DGP measure and the t-GAS model and KL(A) the KL divergence between the DGP and some model A. Then the relative KL divergence plotted in Figure 3 is defined as $1-\text{KL(G)}/\text{KL(A)}$. The closer the curve is to zero, the more similar are the com-

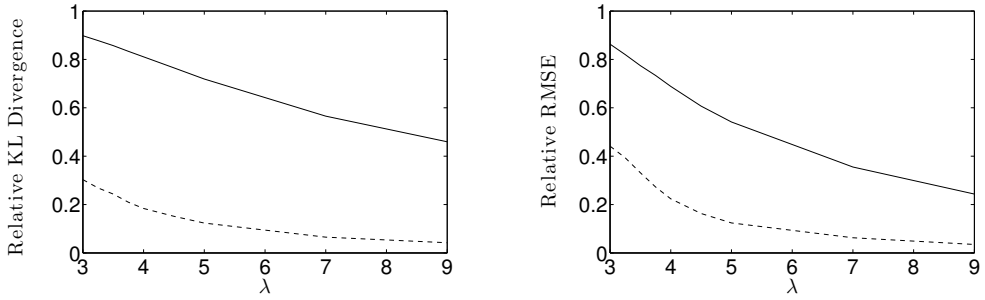

Figure 3: The left panel presents the relative KL divergences for t-GAS relative to GARCH (solid curve) and to t-GARCH (dashed curve). The right panel presents the corresponding relative mean squared error (RMSE) statistics. The results are based on a simulated time series of length $T = 35,000$ for f_t and y_t and for each value of λ using the stochastic volatility DGP with $a = 0.0$, $b = 0.98$, $\sigma_\epsilon = 0.065$.

peting models in terms of their KL divergence with respect to the DGP. If the relative KL divergence approaches one, then this means that the t-GAS model has an arbitrarily small KL divergence compared to the alternative model.

Figure 3 clearly underlines the superiority of the t-GAS update relative to the GARCH and t-GARCH updates. The superiority of t-GAS in comparison to GARCH is not surprising since the GARCH substantially misspecifies the conditional distribution. More interesting is the comparison between the t-GAS and t-GARCH models. In Figure 3 it becomes clearly visible that the score based update yields 10% to 30% decrease in average KL-divergence compared to the t-GARCH models with degrees of freedom in the empirically relevant range of $\lambda \in [3, 6]$. If the conditional distribution is even more fat-tailed, the improvements amount to almost 50%.

The same conclusions are obtained from the relative root mean squared errors (RMSEs) which are displayed in the right-hand panel of Figure 3. The RMSEs based on $(\tilde{f}_t - f_t)^2$ can be computed since the f_ts are observed in our current simulation setting. The relative improvements of the score based t-GAS steps vis-à-vis the GARCH and the t-GARCH steps are substantial. We find improvements with respect to the t-GARCH model of approximately 10% for $\lambda = 6$ and over 40% for $\lambda = 3$.

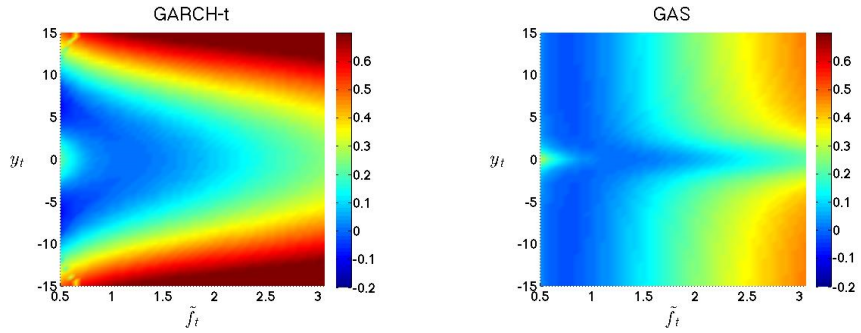

Figure 4: The RKL variation and RKL optimality regions (regions with negative RKL variation) from Definition 1 for t-GARCH (left-panel) and t-GAS (right-panel) updates. The conditional density of y_t given f_t used in these pictures is the standard Student's t with $\lambda = 3$ degrees of freedom. The regions are plotted for a given true $f_t \approx 1.2$.

4.4 Realized Kullback-Leibler variation

Figure 4 presents the realized KL (RKL) variation from Definition 1 for the t-GARCH and t-GAS models. The RKL optimality regions correspond to the areas with negative KL variation. In order to highlight the differences, the RKL variation is plotted using the pseudo-true parameters of the two observation driven models for the DGP with $\lambda = 3$. We plot the regions for a given true $f_t \approx 1.2$ for a range of y_t and $\tilde{f}_t$ that contains almost 99% of the mass of their respective simulated paths.

Figure 4 shows that if $\tilde{f}_t$ is below the true value $f_t \approx 1.2$, then most of th t-GARCH and t-GAS updates improve the KL divergence for large values of $|y_t|$. Large values of $|y_t|$ force $\tilde{f}_t$ to be updated upwards. By contrast, if $\tilde{f}_t$ is larger than the true value $f_t \approx 1.2$, then only small values of $|y_t|$ make $\tilde{f}_t$ converge downwards to its mean due to autoregressive dynamics of GARCH and GAS models. However, Figure 4 also shows that the t-GARCH is much more sensitive to large values of $|y_t|$. Due to the form of the parameter update, large values of $|y_t|$ lead to an increase of $\tilde{f}_t$ such that its approximation to the true density of f_t deteriorates significantly. For the GAS model, on the other hand, the effect of y_t on the update of $\tilde{f}_t$ is bounded. Hence, the GAS update achieves a much larger region where the update reduces the KL divergence.

4.5 Conditionally expected Kullback-Leibler variation

Figure 5 presents the conditionally expected KL (CKL) variation for the t-GARCH and t-GAS models. The CKL optimality regions correspond to the areas where the CKL variation is negative. For both models, the regions close to the 45 degree line where $f_t \approx \tilde{f}_t$ are the most problematic. In those regions the signal is less informative. In particular, if $\tilde{f}_t$ is substantially lower (higher) than f_t, the observed y_t is likely to be informative that the filtered $\tilde{f}_t$ needs to be updated upwards (downwards). In this case, the KL variation will be negative with high probability. But when $\tilde{f}_t$ is close to f_t, the randomness of y_t can easily lead f_t to be updated incorrectly. For example, for known λ and in the extreme case of $\tilde{f}_t = f_t$, the update for $\tilde{f}_t$ can only keep the same KL divergence to the true conditional density for a given value of y_t if the square (GARCH) or score (GAS) compensates for the change due to the autoregressive term. Such an exact compensation occurs with probability zero.

After a closer inspection of Figure 5, we may conclude that the GAS model behaves more favorable than the t-GARCH model: (i) the regions of positive KL variation (in orange and red, around the diagonal) are considerable smaller for GAS, (ii) the regions of negative CKL variation (in green and blue, outside the diagonal) are considerably wider for the GAS model. Whenever $\tilde{f}_t$ is substantially different from f_t, the GAS model presents stronger expected reductions in KL divergence.

5 Conclusions

We have provided an information theoretic foundation for the use of the score of the conditional model density in updating the time-varying parameters in observation driven time varying parameter models. Such score driven models are known as generalized autoregressive score models and have been applied successfully for empirical studies in economics and finance. We have shown that updates based on the score minimize the local Kullback-Leibler divergence between the

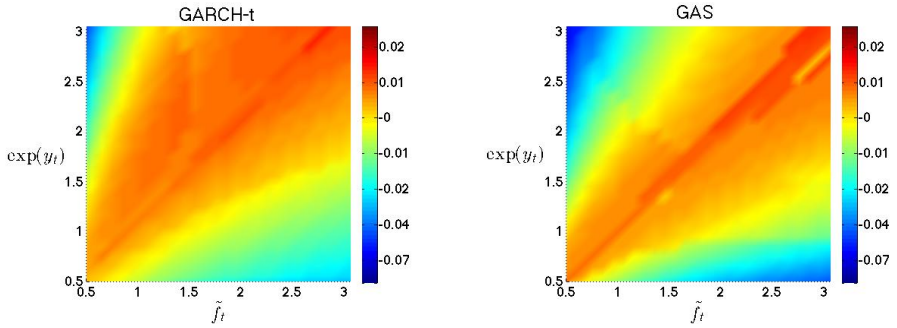

Figure 5: The conditionally expected KL (CKL) variation for t-GARCH (left-panel) and t-GAS (right-panel) updates. The conditional density of y_t given f_t is the standard Student's t with $\lambda = 3$ degrees of freedom.

(unknown) true conditional data density and the model implied conditional density. We have also established conditions under which the updates are optimal in a non-local sense. The numerical results presented in our simulation study on volatility models revealed that updates based on the score reduce the Kullback-Leibler divergence more frequently than existing models. The numerical results also showed that score updates in the volatility context not only reduce local, but also global versions of Kullback-Leibler divergence.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, Armenian SSR*, pp. 267–281.

Andres, P. (2014). Computation of maximum likelihood estimates for score driven models for positive valued observations. *Computational Statistics and Data Analysis*, forthcoming.

Blasques, F., S. J. Koopman, and A. Lucas (2012). Stationarity and ergodicity of univariate generalized autoregressive score processes. *Discussion Paper, Tinbergen Institute* (12-059).

Blasques, F., S. J. Koopman, and A. Lucas (2014). Maximum likelihood estimation for generalized autoregressive score models. *Discussion Paper, Tinbergen Institute* (14-029/III).

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*(3), 307–327.

Cover, M. and J. Thomas (1991). *Elements of Information Theory*. Wiley, New York.

Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics 8*, 93–115.

Creal, D., S. J. Koopman, and A. Lucas (2008). A general framework for observation driven time-varying parameter models. Discussion Paper 08-108/4, Tinbergen Institute.

Creal, D., S. J. Koopman, and A. Lucas (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business and Economic Statistics 29*(4), 552–563.

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

Creal, D., B. Schwaab, S. J. Koopman, and A. Lucas (2014). Observation driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, forthcoming.

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

De Lira Salvatierra, I. and A. J. Patton (2013). Dynamic copula models and high frequency data. *Duke University Discussion Paper*.

Durbin, J. and S. J. Koopman (1997). Monte Carlo Maximum Likelihood estimation for non-Gaussian State Space Models. *Biometrika 84(3)*, 669–684.

Durbin, J. and S. J. Koopman (2012). *Time series analysis by state space methods*. Oxford University Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations. *Econometrica 50*, 987–1008.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics 17*(5), 425–446.

Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails*. Cambridge University Press.

Harvey, A. C. and A. Luati (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, forthcoming.

Hjort, N. L. and M. C. Jones (1996). Locally parametric nonparametric density estimation. *Annals of Statistics 24*(4), 1433–1854.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Reviews 106*, 620–630.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Kapur, J. N. and H. K. Kesavan (1992). *Entropy Optimization Principles with Applications*. Academic Press, San Diego.

Koopman, S. J., A. Lucas, and M. Scharth (2012). Predicting time-varying parameters with parameter-driven and observation-driven models. *Tinbergen Institute Discussion Papers 12-020/4*.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Lucas, A., B. Schwaab, and X. Zhang (2014). Measuring credit risk in a large banking system: econometric modeling and empirics. *Journal of Business and Economic Statistics*, forthcoming.

Maasoumi, E. (1986). The measurement and decomposition of multidimensional inequality. *Econometrica 54*, 991–997.

Masreliez, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Transactions on Automatic Control 20*(1), 107–110.

Müller, U. K. and P. E. Petalas (2010). Efficient estimation of the parameter path in unstable time series models. *The Review of Economic Studies 77*, 1508–1539.

Nelson, D. B. and D. P. Foster (1994). Asymptotic filtering theory for univariate arch models. *Econometrica*, 1–41.

Oh, D. H. and A. J. Patton (2013). Time-varying systemic risk: Evidence from a dynamic copula model of cds spreads. *Duke University Discussion Paper*.

Shannon, C. E. (1948, july, october). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423, 623–.

Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference 69*, 137–162.

Ullah, A. (2002). Uses of entropy and divergence measures for evaluating econometric approximations and inference. *Journal of Econometrics 107*(1-2), 313–326.

A Appendix: Review of Empirical Applications

The observation driven time-varying parameter models based on the score of the predictive likelihood function of Creal, Koopman, and Lucas (2008, 2013) and Harvey (2013) have been successfully applied in various empirical studies. Creal, Koopman, and Lucas (2011) adopt the score framework to specify a multivariate heavy-tailed model for time-varying volatilities and correlations. Extensions to multivariate skewed distributions have been further explored by Lucas, Schwaab, and Zhang (2014). Harvey (2013), Koopman, Lucas, and Scharth (2012), Andres (2014), and Harvey and Luati (2014) propose different dynamic location and scale models that are specific applications of the predictive score framework, while new dynamic copula models have been investigated by Oh and Patton (2013) and De Lira Salvatierra and Patton (2013). In an extensive empirical study, Creal, Schwaab, Koopman, and Lucas (2014) show that the generalized mixed measurement dynamic factor models for large unbalanced panels and mixtures of discrete and continuous random variables can be analyzed jointly within the score framework.

B Appendix: Proofs

Proof of Proposition 1. The line of proof presented below set the stage for all subsequent proofs. Let $Y_{\delta_y}(y_t)$ be as defined in (8). By a repeated application of the mean value theorem to $\tilde{p}(y|\tilde{f}_{t+1};\boldsymbol{\theta})$ and $\tilde{\nabla}_t(\tilde{f}^*_{t+1}, y_t;\boldsymbol{\theta})$, and using the form of the Newton-GAS update $\tilde{f}_{t+1} - \tilde{f}_t = \alpha S(\tilde{f}_t;\boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta})$, we obtain CKL

optimality if

$$
\begin{aligned}
& \int_{Y_{\delta_y}(y_t)} p(y|f_t) \ln \frac{\tilde{p}(y|\tilde{f}_t;\boldsymbol{\theta})}{\tilde{p}(y|\tilde{f}_{t+1};\boldsymbol{\theta})} \mathrm{d}y \\
= \quad & -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \frac{\partial \ln \tilde{p}(y|\tilde{f}^*_{t+1};\boldsymbol{\theta})}{\partial f} (\tilde{f}_{t+1} - \tilde{f}_t) \mathrm{d}y \\
= \quad & -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \tilde{\nabla}(\tilde{f}^*_{t+1}, y; \boldsymbol{\theta}) \; \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta}) \mathrm{d}y \\
= \quad & -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \Big(\tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta})\Big)^2 \mathrm{d}y \qquad (16) \\
& -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta}) \frac{\partial \tilde{\nabla}(\tilde{f}^{**}_{t+1}, y^{**}_t;\boldsymbol{\theta})}{\partial y} (y_t - y) \mathrm{d}y \\
& -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta}) \frac{\partial \tilde{\nabla}(\tilde{f}^{**}_{t+1}, y^{**}_t;\boldsymbol{\theta})}{\partial f} (\tilde{f}^*_{t+1} - \tilde{f}_t) \mathrm{d}y < 0, \\
=: \quad & -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \Big(\tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta})\Big)^2 \mathrm{d}y + A(\delta_f, \delta_y) + B(\delta_f, \delta_y), \qquad (17)
\end{aligned}
$$

where $\tilde{f}^*_{t+1}$ is a point between $\tilde{f}_{t+1}$ and $\tilde{f}_t$, $\tilde{f}^{**}_{t+1}$ is a point between $\tilde{f}^*_{t+1}$ and $\tilde{f}_t$, y^{**}_t is a point between y_t and y, and $A(\delta_f, \delta_y)$ and $B(\delta_f, \delta_y)$ in (17) are equal to the second and third term of (16), respectively. From Assumptions 1 and 2 we obtain $\alpha S(\tilde{f}_t;\boldsymbol{\theta}) \Big(\tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta})\Big)^2 > 0$ almost surely, such that for every $\tilde{f}_t$ and p_t, $\exists \gamma < 0$ such that

$$
-\int_{Y_{\delta_y}(y_t)} p(y|f_t) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \Big(\tilde{\nabla}(\tilde{f}_t, y_t;\boldsymbol{\theta})\Big)^2 \mathrm{d}y \leq \gamma < 0.
$$

The desired result now follows by noting that second and third term in (16) can be made arbitrarily small compared to the first term due to the differentiability of the score and the compactness of $Y_{\delta_y}(y_t)$; see working paper for more details.

The proof for local CKL-optimality follows immediately by a similar argument using the assumption that $\tilde{f}_{t+1}$ is a continuous random variable with a density. □

Proof of Proposition 2. Let

$$
\tilde{f}_{t+1} - \tilde{f}_t = \phi(\tilde{f}_t, y_t;\boldsymbol{\theta}) - \phi(\tilde{f}_{t-1}, y_{t-1};\boldsymbol{\theta}) = \Delta\phi(\tilde{f}_t, y_t;\boldsymbol{\theta}).
$$

We follow the same line of proof as for Proposition 1. To prove the 'if' part, we write the local RKL variation for any given p_t as

$$\begin{aligned} -\int_{Y_{\delta_y}(y_t)} p(y|f_t) \frac{\partial \ln \tilde{p}(y|\tilde{f}^*_{t+1};\boldsymbol{\theta})}{\partial f}(\tilde{f}_{t+1}-\tilde{f}_t)\mathrm{d}y \\ = -\int_{Y_{\delta_y}(y_t)} p(y|f_t)\tilde{\nabla}(\tilde{f}^*_{t+1},y;\boldsymbol{\theta})\Delta\phi(\tilde{f}_t,y_t;\boldsymbol{\theta})\mathrm{d}y. \end{aligned} \tag{18}$$

Using the definition of a score-equivalent update and the same argument as in the proof of Proposition 1, we have, for a sufficiently small δ_y,

$$\mathrm{sign}(\tilde{\nabla}(\tilde{f}^*_{t+1},y;\boldsymbol{\theta})) = \mathrm{sign}(\Delta\phi(\tilde{f}_t,y_t;\boldsymbol{\theta})) \quad \forall\ (f,y) \in F_{\delta_f}(\tilde{f}_t)\times Y_{\delta_y}(y_t),$$

and hence

$$\int_{Y_{\delta_y}(y_t)} p(y|f_t)\tilde{\nabla}(\tilde{f}^*_{t+1},y;\boldsymbol{\theta})\Delta\phi(\tilde{f}_t,y_t;\boldsymbol{\theta})\mathrm{d}y > 0.$$

It implies that the local RKL variation in (18) is strictly negative under the regularity conditions of Assumptions 1 and 2. A similar argument holds for CKL variation by taking a subsequent expectation over $\tilde{f}_{t+1}$ given $\tilde{f}_t$ and f_t.

To prove the 'only if' part, suppose that the update $\tilde{f}_{t+1} = \phi(\tilde{f}_t, y_t;\boldsymbol{\theta})$ is not score-equivalent. Then, by Assumption 1, there must exists an open set $FY \subseteq \mathcal{F}\times\mathbb{R}$ such that $\mathrm{sign}(\tilde{\nabla}(f,y;\boldsymbol{\theta})) \neq \mathrm{sign}(\Delta\phi(f,y;\boldsymbol{\theta}))$ for all $(f,y)\in FY$. This implies in turn that for sufficiently small δ_y we get $\tilde{\nabla}(\tilde{f}^*_{t+1},y;\boldsymbol{\theta})\Delta\phi(\tilde{f}_t,y_t;\boldsymbol{\theta}) < 0$ for all $(f,y)\in FY$. Hence, by Assumption 1, $\exists\delta_y > 0$ such that

$$\int_{Y_{\delta_y}(y_t)} p(y|f_t)\tilde{\nabla}(\tilde{f}^*_{t+1},y;\boldsymbol{\theta})\Delta\phi(\tilde{f}_t,y_t;\boldsymbol{\theta})\mathrm{d}y < 0.$$

Hence, a non-score equivalent update is not RKL optimal regardless of p_t. By Assumption 1, the result extends immediately to CKL optimality. □

Proof of Proposition 3. As in the proof of Proposition 1, we require

$$\begin{aligned}
\Delta_{t|t} &= -\int_Y p(y|f_t)\tilde{\nabla}(\tilde{f}_t^*, y; \boldsymbol{\theta})(\tilde{f}_{t+1} - \tilde{f}_t)\mathrm{d}y \\
&= -\int_Y p(y|f_t)\tilde{\nabla}(y|\tilde{f}_t^*)(\omega + \alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(f_t, y; \boldsymbol{\theta}) + (\beta - 1)\tilde{f}_t)\mathrm{d}y \\
&= -\int_Y p(y|f_t)\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})(\omega + (\beta - 1)\tilde{f}_t)\mathrm{d}y \qquad (19) \\
&\quad -\int_Y p(y|f_t)\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})^2\mathrm{d}y + A(\delta_y, \delta_f) \;<\; 0
\end{aligned}$$

where $A(\delta_y, \delta_f)$ is an appropriate remainder term as in the proof of Proposition 1, which can be made arbitrarily small by selecting small enough values for (δ_y, δ_f).

The second term in (19) is surely strictly negative, whereas the first term may not be. As a result, for small enough (δ_y, δ_f) we obtain the desired inequality if

$$\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t))^2 > |\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|\,|\omega + (\beta - 1)\tilde{f}_t| \;\Leftrightarrow\; \alpha > \frac{|\omega + (\beta - 1)\tilde{f}_t|}{S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|}.$$

The proof for local CKL optimality follows by a similar argument and by taking additional local expectations with respect to $\tilde{f}_{t+1}$ given $\tilde{f}_t$ and f_t. □

Proof of Proposition 4. From the proof of Proposition 1, $\Delta_{t|t} < 0$ holds if

$$\begin{aligned}
&- \int_{Y_{\delta_y}(y_t)} p(y|f_t)\alpha S(\tilde{f}_t; \boldsymbol{\theta})\Big(\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})\Big)^2\mathrm{d}y \\
&- \int_{Y_{\delta_y}(y_t)} p(y|f_t)\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})\frac{\partial\tilde{\nabla}(\tilde{f}_{t+1}^{**}, y_t^{**}; \boldsymbol{\theta})}{\partial y}(y_t - y)\mathrm{d}y \\
&- \int_{Y_{\delta_y}(y_t)} p(y|f_t)\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})\frac{\partial\tilde{\nabla}(\tilde{f}_{t+1}^{**}, y_t^{**}; \boldsymbol{\theta})}{\partial f}(\tilde{f}_{t+1}^* - \tilde{f}_t)\mathrm{d}y < 0.
\end{aligned}$$

The first term is strictly positive, but the other two may be either positive or negative. By norm subadditivity, we obtain that the inequality is satisfied if

$$\begin{aligned}
\alpha S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|^2 &> \alpha S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|\,|\zeta(\delta_f, \delta_y)|\,\delta_y \\
&+ \alpha S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})|\,|\xi_f(\delta_f, \delta_y)|\,\delta_f \;\Leftrightarrow \qquad (20) \\
|\tilde{\nabla}(\tilde{f}_t, y_t; \boldsymbol{\theta})| &> \xi(\delta_f, \delta_y) \times \delta_f + \zeta(\delta_f, \delta_y) \times \delta_y = \eta(\delta_f, \delta_y).
\end{aligned}$$

A similar argument holds for the CKL optimality. □

Proof of Proposition 5. From the proofs of Proposition 1 and 3, we have

$$\begin{aligned}
\Delta_{t|t} = &- \int_Y p(y|f_t)\tilde{\nabla}(y_t|\tilde{f}_t)(\omega + (\beta-1)\tilde{f}_t)\mathrm{d}y \\
&- \int_Y p(y|f_t)\frac{\partial\tilde{\nabla}(y_t^{***},\tilde{f}_t^{***})}{\partial y}(\omega + (1-\beta)\tilde{f}_t)(y-y_t)\mathrm{d}y \\
&- \int_Y p(y|f_t)\frac{\partial\tilde{\nabla}(y_t^{***},\tilde{f}_t^{***})}{\partial \tilde{f}}(\omega + (\beta-1)\tilde{f}_t)(\tilde{f}_t^{*}-\tilde{f}_t)\mathrm{d}y \\
&- \int_Y p(y|f_t)\alpha S(\tilde{f}_t;\boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t)^2\mathrm{d}y \\
&- \int_Y p(y|f_t)\alpha S(\tilde{f}_t;\boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t)\frac{\partial\tilde{\nabla}(y_t^{**},\tilde{f}_t^{**})}{\partial y}(y-y_t)\mathrm{d}y \\
&- \int_Y p(y|f_t)\alpha S(\tilde{f}_t;\boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t)\frac{\partial\tilde{\nabla}(y_t^{**},\tilde{f}_t^{**})}{\partial f}(\tilde{f}_t^{*}-\tilde{f}_t)\mathrm{d}y,
\end{aligned}$$

with the fourth term strictly positive and the remaining terms possibly taking positive or negative values. As such, we obtain the desired result once more by ensuring the fourth term is larger in absolute value than the sum of the remaining terms. By norm sub-additivity, this is implied by

$$\begin{aligned}
\alpha S(\tilde{f}_t;\boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t)^2 > &|\tilde{\nabla}(y_t|\tilde{f}_t)||\omega + (\beta-1)\tilde{f}_t| + \zeta_{\delta_f,\delta_y}(\tilde{f}_t,y_t;\boldsymbol{\theta})|\omega + (1-\beta)\tilde{f}_t|\delta_y \\
&+ \xi_{\delta_f,\delta_y}(\tilde{f}_t,y_t;\boldsymbol{\theta})|\omega + (\beta-1)\tilde{f}_t|\delta_f \\
&+ \alpha S(\tilde{f}_t;\boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t)\Big[\zeta_{\delta_f,\delta_y}(\tilde{f}_t,y_t;\boldsymbol{\theta})\delta_y + \xi_{\delta_f,\delta_y}(\tilde{f}_t,y_t;\boldsymbol{\theta})\delta_f\Big],
\end{aligned}$$

which delivers the desired result. A similar argument holds for CKL optimality.

□