ALMA MATER STUDIORUM - UNIVERSITÁ DI BOLOGNA

DOTTORATO DI RICERCA IN: Scienze Statistiche

Ciclo: XXXIII

Settore Concorsuale: 13/D1 Statistica

Settore Scientifico Disciplinare: SECS-S/01 Statistica

On Optimality of Score Driven Models

Presentata da:
Christopher Sacha Aristide Lauria

Supervisore
Prof. Alessandra Luati

Coordinatore
Prof. Alessandra Luati

Esame finale anno 2021

“Noble deeds and hot baths are the best cures for depression.”
— Dodie Smith

“Create all the happiness you are able to create; remove all the misery you are able to remove.”
— Jeremy Bentham

Statutory declaration

I hereby declare that I have developed and written the following PhD Thesis completely by myself. Wherever contributions of others are involved, every effort has been made to indicate this clearly through references to the Bibliography and acknowledgements.
Bologna,

Christopher Sacha Aristide Lauria

Contents

Abstract

The contribution of this thesis consists in proving that score driven models possess a novel, intuitive, high dimensional and global optimality criterion, called Conditional Expected Variation optimality that formalizes the following words from Creal et al. (2013): " The use of the score is intuitive. It defines a steepest ascent direction for improving the model's local fit in terms of the likelihood or density at time t given the current position of the parameter. This provides the natural direction for updating the parameter" .

Indeed, the fact that the score defines a steepest ascent direction is crucial in deriving the results and for the proposed optimality criterion to hold. To prove the aforementioned property, a point of contact between the econometric literature and the time varying optimization literature will be established. As a matter of fact, the Conditional Expected Variation optimality can be naturally viewed as a generalization of the monotonicity property of the gradient descent scheme. A number of implications on the specification of score driven models are analyzed and discussed, even in the case of model misspecification.

Key words and phrases: Time varying parameters; Score functions; Gradient descent.

Acknowledgements

First I would like to thank my supervisor, Professor Alessandra Luati, for correcting my innumerable writing mistakes, giving me the opportunity to do research, and allowing me to explore my ideas to the fullest.

My thanks extend also to Professor Siem Jan Koopman, that welcomed me with open arms at VU university, during my stay abroad, and greatly encouraged me in the pursuit of new results.

I owe my deepest gratitude to Professor Francisco Blasques, for breathing new life into my motivation, for our discussions on the more technical aspects of my research and for giving me a new understanding of the relevant literature. Thank you so much Chico.

My appreciation and thanks to Professor Paolo Gorgi, that expertly delved into the mathematical aspects of my thesis and helped me gain a deeper understanding of the critical issues I was facing.

Last, but not least, I wish to extend my appreciation and praise to both Professor Enrico Bernardi and Professor Alberto Lanconelli; extraordinary educators that constantly inspire me to improve my mathematical knowledge.

Chapter 1

Introduction

Data recorded sequentially over time are typically observed in a plethora of natural and man made phenomena. Examples vary from the amount of millimeters of rain experienced in a certain location to the returns of a financial index. The main assumption behind time series analysis is that past observations of the data contain information about future observations, thus one can utilize past observations to make predictions about the future. To do this, one usually requires a mathematical model. The setting is different than the one of independently and identically distributed (IID) observations, that typically occur in experiments which take place under the same overall conditions. When the data exhibit some dependence we can no longer use the same tools as in the IID case.

Several statistical models have been developed for data that are not indipendent, the most well know being an autoregressive model. The first applications of an autoregressive model date back to Yule (1927) and Slutsky (1927). Walker (1931) will later acknowledged Yule's paper as being an important extension of his ideas on periodicity. Since then a standard procedure for specification, estimation, diagnostic checking and forecasting has been developed by Box et al. (2015) for all the models belonging to the auto-regressive integrated moving average class (ARIMA). ARIMA models have thus enjoyed a great deal of popularity thanks to their simple specification, their fleshed out mathematical theory and the iterative procedure derived by Box et al. (2015). However a limitation they posses is that they only take into account possible linear dependencies in the data, leaving out non-linear relations. In some applications this limitation can be restrictive, so that extensions to larger classes of statistical models, with appropriate properties to handle non-linear relations, turn out to be needed.

In a parametric setting one way to introduce non-linearities is by allowing time variation in some features of the probability distribution that models the data. A simple way to achieve time variation of a parametric probability distribution is by allowing the parameters that characterize the distribution itself to vary through time. This thesis focuses on a class of non-linear models and, in particular, on time-varying parameter models. A valuable point of strenght of time-varying parameter models is an ever-growing body of evidence that the linear regression assumption of fixed parameters often appears invalid even when the datapoints seem indipendent from one another. Indeed, structural changes, specification errors, proxy variables and aggregation are all sources of parameter variation; see Sarris (1973), Belsley (1973), Belsley and Kuti (1973) and Cooley and Prescott (1976).

In Cox et al. (1981) a categorization of time varying parameter models was given by dividing them into two classes: observation driven models and parameter driven models. The former are models where the updating equation is a function of the observations, the latter are models where the dynamic equation is governed by idiosyncratic innovations. For a discussion on strengths and weaknesses of these two classes of models see Koopman et al. (2016). In this thesis we will almost exclusively deal with observation driven models.

In the financial literature, observation driven models have increasingly risen to prominence after Engle, in an effort to explain the observed heteroskedasticity in the variance of financial time series, introduced the observation driven autoregressive conditional heteroskedasticity (ARCH) model Engle (1982). The latter was later generalized by Bollerslev (1986) to the, ubiquitously used, generalized ARCH (GARCH) specification. The GARCH model or, as referred to by Lee and Hansen (1994b), "the workhorse of the industry" has been extensively studied. The asymptotic properties of the quasi maximum likelihood estimator for a GARCH(1,1) have been derived in Lee and Hansen (1994b), the consistency and asymptotic normality for a covariance stationary GARCH(1,1) have been proven in Lumsdaine (1996a), for the general case of a GARCH(p,q), where $p, q \in \mathbb{N}$. Consistency and asymptotic normality of the parameter estimators have been studied in a number of works Berkes et al. (2003), Francq and Zakoian (2004), Alzghool (2017), and the mathematical and statistical properties have been covered in multiple surveys and books, see Bera and Higgins, Shephard (1996), Francq and Zakoian (2010). Despite the great popularity of the GARCH model in applied settings when predicting financial volatility, some shortcomings have been discussed in Nelson (1991), Zivot (2009). Notably, the first order GARCH model does not account for three stylized facts such as the asymmetric distribution of financial returns, the fact that persistence of conditional volatility tends to increase with the sampling frequency and the negative correlation between volatility of current returns and future ones.

Hence, a rapid multiplication of new model specifications that aimed to adress these shortcomings occured. Three of the most prominent extensions are the exponential

GARCH (EGARCH) by Nelson (1991), that aims to capture the negative correlation between volatility of current returns and future ones, the nonlinear asymmetric GARCH (NAGARCH) Engle and Ng (1993) and the Glosten et al. (1993) model (GJR), that both allow for negative returns to increase future volatility by a larger amount than positive returns of the same magnitude. A review is given in Degiannakis and Xekalaki (2004), see also Francq and Zakoian (2010).

Recently, a new class of observation driven models was introduced through the independent work of Harvey (2013) and Creal et al. (2013) with the aim to provide a unified framework. This class of models was formerly known as the Generalized Autoregressive Score (GAS) or Dynamic Conditional Score (DCS) class; now it's commonly referred to as the class of score driven models. The key feature of score driven models is that the dynamic of the time varying parameter is driven by a martingale difference sequence proportional to the score of the conditional likelihood with respect to the parameter of interest.

Although GAS models have been empirically validated multiple times as for example in Creal et al. (2013), Harvey (2013), Harvey and Luati (2014b), Blazsek and Villatoro (2015), Fonseca and Cribari-Neto (2018), Ayala and Blazsek (2018), Catania et al. (2018), Gorgi et al. (2019), Babii et al. (2019) (a repository for score driven papers is available at `http://www.gasmodel.com/gaspapers`) some questions are still open, partly due to the fact that the class of models has been developed relatively recently, partly because some issues related to the use of the score have not been fully developed. Specifically, the role of the score as a driving force in the updating equation has not completely been uncovered. In addition, there are currently no theoretically motivated prescriptions on the proportionality coefficient that multiplies the score in the dynamic equation.

In the direction of the first question, regarding the use of the derivative of the loglikelihood in the dynamic equation, a novel property related to the use of the score was given in Blasques et al. (2015), Blasques et al. (2018) where it was proven that (under a sign condition) first order GAS models satisfy, locally, an information-theoretic optimality criterion based on the Kullback Leibler (KL) variation. The setting used in Blasques et al. (2015) provides an original framework for deriving further results on observation driven models that we will adopt for our analysis.

As a matter of fact, the use of the score as a driving force in the updating equation for the time varying parameter is very intuitive: in the words of Creal et al. (2013) "The use of the score for updating the parameter is intuitive. It defines a steepest ascent direction for improving the model's local fit in terms of the likelihood or density at time t given the current position of the parameter. This provides the natural direction for updating the parameter.”. However, the same intuition does not find formal evidence in the current literature.

This thesis formalizes this intuition and proves that indeed the score is a natural choice as a driving force of the updating equation in the context of the time varying optimization theory. More precisely we consider a time varying optimization problem, where the sequence of objective functions is given by the model log-likelihoods at each time, and resort to the stochastic gradient descent literature to solve it. In doing so we prove, under some assumptions on the model density, that each iteration of a specific score driven model moves closer, in Euclidean distance and conditional expectation, to the (pseudo)-true time varying parameter. We will refer to this property, that is a natural consequence of the monotonicity of the updates of a gradient descent scheme, as to the optimal conditional expected variation (CEV).

Unlike Blasques et al. (2015), where the optimality criterion is based on the Kullback Leibler divergence between the model density and the true density, the conditional expected variation regards Euclidean distances in the parameter space. In addition, conditional expected variation optimality holds on the whole parameter space, and is trivially extended to the case in which the time varying parameter is multi dimensional. Initially the assumptions on the model density, used to derive the original results the thesis has to offer, are quite restrictive. Taking this fact into account, we then provide a series of extensions to a more general setting. The choice of the scaling function present in the score driven model specification will be crucial for the extensions to hold.

The use of the score in the dynamic equation will be essential to obtain the results. Most of the derivations of the theorems will proceed from the theory of optimization in conjunction with the framework developed in Blasques et all Blasques et al. (2015).

In summary, there are three main contributions of the thesis:

- The specification of a time varying optimization problem formalizes the fact that the score is a natural choice as a driving mechanism for the dynamic equation.
- We give explicit conditions, that can be checked in an applied setting, for score driven models to achieve the aforementioned CEV property.
- We obtain constraints on the scaling function, present in the score driven specification given in Creal et al. (2013), and retroactively formally motivate some of the choices made in the literature when defining this function.

The thesis is structured as follows: First, an introduction to the theory behind gradient descent is provided in chapter 2, to make the framework, that leads to the proof of CEV optimality, clear. The standard case of a static objective function is considered under the assumptions of strong convexity and Lipschitz continuous gradient, which allow one to obtain simple proofs of the convergence of the gradient descent algorithm. Secondly, the static case is generalized in two different directions. On one hand the stochastic case, where the well known results on almost sure convergence and convergence in expectation are proved. On the other hand, to the case of a time varying objective function that changes minima at each time t, here some of the more modern literature is displayed. Prediction-correction methods will be introduced along with the complexity given by the time dependency of the objective function. Finally some original convergence results, regarding the the time varying objective function, are proven by imposing a novel assumption on the dynamics of the minima through time. We present the optimization literature as separate from the econometric theory surrounding score driven models, the point of contact will be made explicit in section 4.2.

In chapter 3 observation driven models will be presented with a focus on the class of score driven models. The full specification of score driven models is given in its entirety. Some of the more notorious examples of score driven specifications are presented; among which the Beta-t-EGARCH model. Here the arbitrariness of the function S_t characterizing the score driven model specification will be highlighted since it will be later object of analysis. Then the latest results on consistency and asymptotic normality of the maximum likelihood estimators of the parameters are provided. We conclude with an introduction to the concept of model invertibility and explain its importance in relation to maximum likelihood estimators.

Chapter 4 reviews the current literature on optimality properties of score driven models. Definitions of key concepts like realized Kullback-Leibler optimality and Newton-score update are stated and the main theorem of Blasques et al. (2015) is proved and discussed.

In section 4.2, that contains most of the original contributions of the thesis, we define optimality in conditional expected variation (CEV) 4.2.1 and prove that the Newton-GAS update is, in fact, CEV optimal. The conditional expected variation property, although inspired by the previous literature on optimality of score driven models, will have the novel characteristics of being global, intuitive and applicable to the high dimensional case.

To prove CEV optimality of the Newton-GAS update we define a time varying stochastic optimization problem thus providing a novel point of contact between the optimization literature and the econometric one. The assumptions used to prove CEV optimality are then gradually relaxed, through a series of propositions, until they encompass some of the more popular choices of model density. Subsequently, we will give some extensions of CEV optimality, under novel assumptions on the data generating process. Here we will depart from the Newton-score update and prove propositions where a larger class of score driven model specifications are utilized. Moreover a connection between CEV optimality and the invertibility property will be spelled out.

Finally conclusions will be drawn and the possibilities of new models based on the developed theoretical framework will be touched upon.

Propositions without citation are to be considered original work as far as the author knows.

Chapter 2

Gradient Descent

The optimization technique known as gradient descent was first introduced by Louis Augustin Cauchy (1847). It is a first-order iterative optimization algorithm for finding the local minimum of differentiable functions. This procedure, that aims to minimize the differentiable objective function f, by iteratively moving in its direction of steepest descent, can be described as follows. Given a starting point x_0, the gradient descent algorithm updates the subsequent values $x_1, x_2, \ldots$ according to the recursive equation

$$x_{t+1} = x_t - \alpha \nabla f(x_t) \tag{2.1}$$

where $\alpha \in \mathbb{R}$ is a hyper-parameter (called the learning rate) and $\nabla f(x_t)$ is the gradient of f calculated at the point x_t.

The resulting algorithm implied by Cauchy's work started to gain major applied importance with the advent of electronic calculators. Later the algorithm was thoroughly studied and generalized in a number of different ways, Nesterov (2014) and Qian (1999). A comparative analysis of various extensions in terms of convergence speed and accuracy can be found in Dogo et al. (2018).

The reason why, in the twenty-first century, gradient descent has become one of the most popular algorithms to perform optimization is that, in applied problems, functions are often high dimensional and complex (Koch et al. (1999), Bates et al. (1996), Krizhevsky et al. (2012)) so finding a solution analitically is out of the question. Gradient descent offers a simple scheme that requires only the gradient of the objective function to be implemented. In particular, machine learning theory has benefitted greatly from gradient descent as training deep neural networks requires the minimization of a high dimensional and complex objective cost function Amari (1993).

In the following section, some preliminary results on convergence of the gradient descent algorithm will be proven. The subsequent notation will be used:

- Given a generic matrix A we define $||A|| = \max_{||x||=1} ||Ax||$.

- Given A, B symmetric matrices, $A \geq B$ means that the matrix $A - B$ is positive semi-definite.

2.1 Main assumptions

The theoretical framework that deals with gradient descent has been developed and popularized throughout the 20th century and can be presented through two main assumptions on the objective function: strong convexity and a Lipschitz continous gradient. Although we choose to work with these two assumptions on the objective function f, as in Bottou et al. (2016) and Polyak (1987), these hypotheses are not strictly required for the proper functioning of the gradient descent algorithm. Multiple papers on optimization methods have made a great deal of effort in weakening these assumptions. Nonetheless the theory is cleaner to present when they hold. The motivation to weaken the assumptions stems from the objective functions utilized in applied settings, one for all, the ones resulting from deep neural networks. In these cases the loss function on which to perform gradient descent is, usually, neither convex nor does it have a Lipschitz continuous gradient, nonetheless these cases are of the utmost importance if one wishes to train a neural network without getting stuck in local minima.
In this thesis we will weaken the standard assumptions only when necessary. Since keeping the theory as simple as possible will be of pedagogical value for explaining the subsequent connection with score driven models.

Assumption 2.0.1. Lipschitz Continuity of the Gradient

Let the objective function $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable with gradient function $\nabla f : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ that is Lipschitz continuous of constant $\mathrm{L} > 0$, i.e.

$$||\nabla f(x) - \nabla f(y)|| \leq \mathrm{L}||x - y|| \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 2.0.2. Strong Convexity

Let the objective function $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and strongly convex, i.e. there exists a constant $\ell > 0$ such that

$$f(x) \geq f(y) + \nabla f(y)(x - y) + \frac{1}{2}\ell||x - y||^2 \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

Note that strong convexity implies the objective function f has a unique minimum at

x^*. Comparing the definition of strong convexity with the first order convexity condition for a differentiable function f, i.e.,

$$f(x) \geq f(y) + \nabla f(y)(x-y)$$

we immediately recognize that strong convexity implies convexity and strict convexity. In the following we give a sequence of standard propositions that will acquaintance the reader with the optimization theory. Based on these results we will prove convergence of the gradient descent scheme and the subsequent extensions of the theory that will be used when dealing with observation driven models. All proofs of this chapter are deferred to the appendix A .

2.1 Preliminaries

All the statements in this section are well known, in general here we will follow Polyak (1987) with some small deviations.

First, we give a lemma that shows that Lipschitz continuity of the gradient and strong convexity are deeply related to one another.

Lemma 2.1.1. Polyak (1987)
A convex function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies assumption 2.0.1 if and only if

$$f(x) \leq f(y) + \nabla f(y)(x-y) + \frac{1}{2}L||x-y||^2 \quad \forall x, y \in \mathbb{R}^d.$$

Comparing this inequality with the one in 2.0.2 we notice the symmetry between assumptions 2.0.1 and 2.0.2. Indeed these definitions are dual of one another: a function is strongly convex with respect to some norm if and only if its Fenchel conjugate function is Lipschitz with respect to its dual norm, see Kakade and Shalev-Shwartz (2009) for details.

The next lemma will be very convenient for checking when a doubly differentiable function obeys both assumption 2.0.1 and 2.0.2.

Lemma 2.1.2. Polyak (1987)
A twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies assumptions 2.0.1 and 2.0.2 if and only if

$$I_{d\times d}\ell \leq \nabla^2 f(x) \leq I_{d\times d}\mathrm{L} \quad \forall x \in \mathbb{R}^d$$

where $\ell > 0, \mathrm{L} > 0$ and $I_{d\times d}$ is the identity matrix of dimension d.

Thus a twice differentiable real valued function is Lipschitz continuous and strongly convex if and only if its second derivative is bounded by positive quantities both from above and from below. In our setting we have that assumption 2.0.1 implies the right hand side inequality $\nabla^2 f(x) \leq I_{d\times d}\mathrm{L}$ for all x. Instead, the strong convexity assumption 2.0.2 is equivalent to assuming that $\nabla^2 f(x) \geq I_{d\times d}\ell$ for all x.
The bounds on the Hessian make us hopeful in thinking we might be able to recover a surrogate of the mean value theorem for the gradient function, indeed this is possible.

Lemma 2.1.3. Polyak (1987)
Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy both assumption 2.0.1 and 2.0.2, then

$$\nabla f(x) - \nabla f(y) = A(x-y) \quad \forall x, y \in \mathbb{R}^d$$

where $I_{d\times d}\ell \leq A \leq I_{d\times d}\mathrm{L}$.

This result is not as trivial as it may appear since the gradient is a vector valued function for which no exact analog of the mean value theorem can be formulated, indeed to prove it the fundamental theorem of calculus will be used.
The next simple lemma will be used when proving convergence of the stochastic gradient descent scheme

Lemma 2.1.4. Polyak (1987)
Let a function $f : \mathbb{R}^d \to \mathbb{R}$ satisfy assumption 2.0.2 then

$$\nabla f(x)(x - x^*) \geq \ell ||x - x^*||^2.$$

Another relevant result is the Polyak-Lojasiewicz (PL) inequality that is implied by assumption 2.0.2

Proposition 2.1.5. Polyak-Lojasiewicz Polyak (1987)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be under assumption 2.0.2 then

$$\frac{1}{2\ell}||\nabla f(x)||^2 \geq f(x) - f(x^*) \geq \frac{\ell}{2}||x - x^*||^2 \qquad \forall x \in \mathbb{R}^d. \tag{2.2}$$

The left hand side inequality (reffered to as the PL inequality) states that the gradient of a strongly convex function grows more than quadratically as we move away from the optimal function value. If one were to take the inequality 2.2 as a simple condition on

the objective function, it would be sufficient to show a global linear convergence rate for gradient descent Karimi et al. (2016).
There is an analog of the Polyak-Lojasiewicz inequality in the case of a Lipschitz continuous gradient

Lemma 2.1.6. Polyak (1987)
Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy assumption 2.0.1 and let it have a global minimum at x^*. Then

$$\frac{1}{2L}||\nabla f(x)||^2 \leq f(x) - f(x^*) \leq \frac{L}{2}||x - x^*||^2 \qquad \forall x \in \mathbb{R}^d.$$

Sometimes we will also consider functions with Lipschitz continuous gradient that are only convex and not strongly convex; for these functions the co-coercivity of the gradient will be used to prove convergence of the gradient descent scheme

Lemma 2.1.7. Co-coercivity Polyak (1987)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and satisfy assumption 2.0.1 then

$$\left(\nabla f(x) - \nabla f(y)\right)^T (x - y) \geq \frac{1}{L}||\nabla f(x) - \nabla f(y)|| \quad \forall x, y \in \mathbb{R}^d.$$

We now turn to analyze the asymptotic properties of the gradient descent scheme 2.1.

2.2 Convergence of Gradient Descent

In this section we show convergence of the gradient descent scheme and highlight the rate of convergence to the minimum. Which is a well known basic result. We always assume the first step of gradient descent is from an arbitrary point $x_0 \in \mathbb{R}^d$.

Proposition 2.2.1. Polyak (1987)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable and satisfy assumption 2.0.1 and 2.0.2. Then, by running k times the gradient descent update $x_{t+1} = x_t - \alpha\nabla f(x_t)$ with $0 < \alpha \leq 2/\mathrm{L}$, one obtains

$$||x_k - x^*|| \leq q^k||x_0 - x^*||, \qquad q = \max\{|1 - \alpha\ell|, |1 - \alpha\mathrm{L}|\} < 1.$$

The convergence rate of q^k is optimal Polyak (1987). Simply put gradient descent converges, for functions under assumptions 2.0.1 and 2.0.2, regardless the dimension of the space in which the objective function lies and from a computational standpoint it requires only the calculation of the gradient. The next non-standard lemma will be of importance when dealing with objective functions that are not Lipschitz continuous or strongly convex as will be the case in chapter 3.1.

Lemma 2.2.2.
Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable with minimum at x^*.
Let $g : C^2(\mathbb{R}^d) \to C^2(\mathbb{R}^d)$ be such that $I_{d\times d}\ell \leq \nabla(g(\nabla f))(x) \leq I_{d\times d}\mathrm{L}$, for all x, where $0 < \ell \leq \mathrm{L}$, and such that $g(\nabla f)(x^*) = 0$ then running k times the gradient descent update $x_{t+1} = x_t - \alpha g(\nabla f)(x_t)$ for $\alpha \leq 2/\mathrm{L}$ we obtain

$$||x_k - x^*|| \leq q^k ||x_0 - x^*||, \qquad q = \max\{|1-\alpha\ell|, |1-\alpha\mathrm{L}|\} < 1.$$

This simple yet powerful lemma states the obvious: if our objective function f is not Lipschitz in the first derivative, one can use the gradient of a function that is indeed Lipschitz continuous to perform gradient descent successfully, provided that they share the same minimum point x^*. The way the lemma is formulated shows the adjustment one ought to make to the gradient descent update of the original objective function f so that convergence is ensured. It is possible to think of the use of g as a way to replace the original objective function f with $\int_0^x g(\nabla f)(y)\, dy$, to then perform gradient descent on.

A special case of this lemma is given by choosing the function g as $g(\nabla f)(x) := S(x)\nabla f(x)$ for some scaling function $S(x)$, where $I_{d\times d}\ell \leq \nabla(S(x)\nabla f(x)) \leq I_{d\times d}\mathrm{L}$, when the objective function f is strictly convex (but not Lipschitz continuous) and thus has a unique critical point. Then the second condition, namely $S(x^*)\nabla f(x^*) = 0$, is trivially verified without even the knowledge of x^*. Notice that there is no chance of adding an unwanted minimum x_1^* if $S(x_1^*) = 0$ since the imposed bounds guarantee that the function $\int_0^x S(y)\nabla f(y)\, dy$ will be strongly convex and thus have a unique minimum point (in this special case we are also assuming that $\lim_{x\to x^*} S(x)$ does not go to infinity faster than $\nabla f(x) \to 0$). The special case $g(\nabla f)(x) := S(x)\nabla f(x)$ will clarify the role of the scaling function S_t present in the score driven model specification, in particular it will also give conditions on S_t, as will be seen in section 4.2.1.

Example:
Suppose we want to find the minimum point of the function $f(x) = x^4$. Directly utilizing gradient descent would most likely fail (depending on the starting point) since f does not have Lipschitz second derivative. Denote with s the quadratic function $s : x \to x^2$ Choose $g(f) = f/s$ then $g(f')(0) = 0$ and $g'(f')(x) = 4$ so the conditions of lemma 2.2.2 are satisfied and running the iteration $x_{t+1} = x_t - \alpha g(\nabla f)(x_t)$ will indeed give convergence to the minimum.

In this example it is easy to see that we are just performing gradient descent on the function $x \to 2x^2$ that shares a minimum at zero with the original objective function f.

The kind of workaround adopted in lemma 2.2.2 has not got much traction in the recent machine learning literature on gradient descent since the properties of the objective function f are unknown or intractable and the gradient is hardly ever amendable to a modification of this type. On the contrary, in the optimization problem discussed in section 4.2, the objective functions will be chosen by the statistician, making them easily subject to this expedient.

2.3 Stochastic Gradient Descent

Stochastic gradient descent can be viewed as a generalization of the gradient descent algorithm where the gradient is replaced by a random estimate of it. It was first introduced by Robbins and Monro in 1951 and called stochastic approximation method Robbins and Monro (1951). As in the deterministic case, it aims to minimize an objective function f and can be summarized by a single recursive equation starting at an arbitrary point $x_0 \in \mathbb{R}^d$

$$x_{t+1} = x_t - \alpha_t g(X_t, x_t) \tag{2.3}$$

where $\{X_t\}_{t\in\mathbb{N}}$ is a sequence of random variables taking values in $\mathbb{R}^n$, $g : \mathbb{R}^{n+d} \to \mathbb{R}^d$, and the equation $E[g(X_t, x_t)|X^{t-1}] = \nabla f(x_t)$ is satisfies for all t. The hyperparameter $\alpha_t \in \mathbb{R}$ will now be assumed as time varying and the piece of notation $X^{t-1} = \{X_{t-1}, X_{t-2}, \ldots, X_1\}$ has been used. The reason for adding the time dependence to the hyperparameter will be clear when discussing convergence issues, while the choice of the random estimate of the gradient $g(X_t, x_t)$ is in line with the machine learning optimization literature Bottou et al. (2018), Nguyen et al. (2018), Johnson and Zhang (2013a) and will be relevant when dealing with score driven models in section 4.2. Notice also how now x_t is unconditionally a random variable as opposed to the previous case where it was a number.

In practical applications it might be the case that one can only recover an estimate of the gradient of the function she wishes to minimize, but, more often than not, the gradient is purposefully made stochastic. The best example of this practice is given in the machine learning literature where a sample of the data is selected at random to compute the gradient of the cost function at each step of the scheme. This can reduce the computational cost significantly, since it allows one to compute the stochastic gradient with even a single observation (something that is not possible for the deterministic gradient that includes all observations) without hindering the convergence properties of the algorithm, see Bottou et al. (2018) for a comparison between the deterministic gradient descent and the stochastic one in the machine learning setting. Also, due to the randomness present in the update, stochastic gradient descent has a better chance than its deterministic counterpart to not get stuck in local minima, Kleinberg et al. (2018). These facts made stochastic gradient descent the de facto standard for optimizing cost functions of high dimensional machine learning models Du (2019). As its deterministic counterpart, stochastic gradient descent has inspired a number of other optimization algorithms, see Kingma and Ba (2014), Qian (1999), Johnson and Zhang (2013a) just to name a few. Here we will focus only on the original.

2.3.1 Convergence of stochastic gradient descent

Unalike in the deterministic gradient descent case, to prove convergence of stochastic gradient descent it is necessary to determine in first instance which type of stochastic convergence to consider; given the random nature of the updates. Notice also that, due to the randomness in the updates at each step of the scheme, there is no guarantee that one will get closer to the minimum at every step. In this section, we will obtain both convergence in mean square and almost sure convergence.

In both cases the convergence of the stochastic gradient descent scheme is usually proved when the objective function is strongly convex and has Lipschitz continuous gradient, i.e., f satisfies assumptions 2.0.1 and 2.0.2, with an additional assumption on the variance of the stochastic gradient. Often this takes the form of a uniform boundedness condition on the unconditional joint expectation of the unbiased estimator of the gradient $g(X_t, x_t)$, i. e., $E[\|g(X_t, x_t)\|^2] \leq c$ for all t (the expectation notation without subscripts will always indicates that the expectation is with respect to the joint probability distribution of $X_1, X_2, \ldots, X_t$). Unfortunately, this assumption has been found to possibly contradict the strong convexity one, as explained in Nguyen et al. (2018). Thus new assumptions have been proposed in the literature. We will use the following

Assumption 2.3.1. Nguyen et al. (2018)
Let $E[\|g(X_t, x_t)\|^2] \leq M_0(f(x_t) - f(x^*)) + M$ for all t, where $M_0, M > 0$.

This assumption can be derived by requiring that $g(X_t, x_t)$ be Lipschitz for every realization $X_t(\omega)$, i.e, there exists a constant $L_1 > 0$ such that

$$\|g(X_t(\omega), x) - g(X_t(\omega), y)\| \leq L_1 \|x - y\|$$

for all $x, y \in \mathbb{R}^d$ and all t. Indeed assumption 3 is derived as a lemma in Nguyen et al. (2018), see also Johnson and Zhang (2013a). Moreover, with the strong convexity of the objective function f, assumption 2.3.1 implies that $E[\|g(X_t, x_t)\|^2] \leq M_1 \|\nabla f(x_t)\|^2 + M$, where $M_1 > 0$, through an application of proposition 2.1.5. This inequality has been

used as an assumption in Bottou et al. (2018) and Bertsekas and Tsitsiklis (1996).
As in the deterministic case these assumptions on the boundedness of the second moment of the estimate of the gradient are standard but not necessary for convergence. An active area of the optimization research is dedicated to weakening them, see for example Lei et al. (2019).

Proposition 2.3.2. Gower et al. (2019)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and satisfy assumptions 2.0.1 and 2.0.2. Let assumption 2.3.1 hold on the estimates of the gradient. Then by choosing $\alpha_t = c/t$ where c is a constant such that $0 < c < \min\{2/M_0, 1/\ell\}$, we have that

$$E[||x_{t+1} - x^*||^2] \leq \frac{v}{t+1}$$

where $v = \max\{||x_1 - x^*||^2, c^2 M/(\ell c - 1)\}$.

As opposed to its deterministic counterpart, stochastic gradient descent requires an adaptive step-size α_k to obtain convergence in expectation to the minimum x^*. This is due to the variance of the estimate of the gradient. If the variance were to decrease to zero, then a fixed step-size could be enough to recover convergence. Alternatively with a fixed step-size it is possible to prove convergence to a neighborhood of the minimum, see Bottou et al. (2018). Recently, modifications to the stochastic gradient descent scheme that decrease the variance of the estimates of the gradient have been proposed, these are called variance reduction methods, see Johnson and Zhang (2013b) and Reddi et al. (2015).
The stochastic nature of the algorithm naturally leads to also consider almost sure convergence, additional technicalities are required. To prove the almost sure convergence we will use a specific case of the quasimartingale convergence theorem, following Bottou (1999).

Proposition 2.3.3. Métivier (2011)
Let $\{Z_t\}_{t\in\mathbb{N}}$ be a real valued positive stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and adapted to a filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$. Let

$$\sum_{t\geq 1} E[\mathbb{1}_{F_t}(Z_{t+1} - Z_t)] < \infty$$

where $\mathbb{1}$ is the indicator function and $F_t := \{E[Z_{t+1} - Z_t|\mathcal{F}_t] > 0\}$. Then

$$Z_t \xrightarrow{a.s.} Z_\infty \geq 0$$

where Z_∞ is an integrable random variable.

The proof will not be given since it would require some mathematical machinery outside the scope of this thesis, but can be found in Métivier (2011) as a special case of theorem 9.4 taking into consideration proposition 9.5. With this result we are able to prove the almost sure convergence of stochastic gradient descent in the strongly convex and Lipschitz continuous objective gradient case

Proposition 2.3.4. Gladyshev (1965) , Bottou (1999)
Let $f : \mathbb{R}^d$ be continuously differentiable and satisfy assumptions 2.0.1 and 2.0.2. Let assumption 2.3.1 hold on the estimates of the gradient $\{X_t\}_{t\in\mathbb{N}}$. Then choosing $\{\alpha_t\}_{t\in\mathbb{N}}$ such that $\sum_{i=1}^{\infty}\alpha_i = \infty, \sum_{i=1}^{\infty}\alpha_i^2 < \infty$ we have that

$$x_t \xrightarrow{a.s.} x^*$$

The lengthy well known proof is given in the appendix and uses ideas both from Gladyshev (1965) and Bottou (1999).

2.4 Gradient Descent with a Time Varying Objective Function

A possible generalization of the mathematical problem of optimizing an objective function is that of considering an objective function that also varies through time $\{f_t\}_{t\in\mathbb{Z}}$ (this case is also called the non-stationary case, we avoid this terminology since it clashes with the econometric one). We also notice that although we call this the "time varying case" t can theoretically represent any other variable of interest.
For the minimum of a particular objective function f_t, belonging to the sequence, we will use the notation

$$x_t^* := \arg\min_{x\in\mathbb{R}^d} f_t(x) \tag{2.4}$$

Since we do not have a single minimum anymore we need to reformulate our object of study: our goal is now to track the solution of 2.4 for each time t which corresponds to finding the solution trajectory. A way to do this would be to solve 2.4 for each time t. However, solving 2.4 for each sampling time t is not a viable option in most application domains, even for moderate-size problems as already discussed in the case of a static objective function. The requisite computation time for solving each instance of the problem is often times excessive for it to be practical, see Bittanti and Cuzzola (2001). It is also challenging to reasonably bound the time each problem instance will take to be solved, see Boyd and Vandenberghe (2004). Essentially the same problems that leads one to utilize an iterative algorithm like gradient descent as in the single objective function case are still relevant. Thus there is a scientific literature on iterative algorithms that produce a sequence $\{x_t\}_{t\in\mathbb{Z}}$, from an arbitrary starting point $x_0 \in \mathbb{R}^d$, that converges to the solution trajectory. Our focus will be on these methods and in particular we will analyze the scheme subsumed by the equation

$$x_{t+1} = x_t - \alpha \nabla f_t(x_t) \tag{2.5}$$

that is a simple modification of gradient descent where the gradient of the objective function at time t is utilized at the $t+1$-th step of the algorithm.

Gradient descent with a time varying objective function has been studied in multiple papers, see Simonetto et al. (2016), Popkov (2005), Simonetto and Dall'Anese (2017) as examples. The subsequent analysis does not have the objective of being exhaustive rather we develop the necessary machinery that will be used when discussing score driven models.

At first glance one could require that the sequence $\{x_t\}_{t\in\mathbb{Z}}$, computed by the algorithm of choice, will be such that

$$\lim_{t\to\infty} |f_t(x_t) - f_t(x_t^*)| = 0$$

that is a sequence of points x_t that eventually get closer and closer to minimizing the corresponding function f_t through time. Of course a sequence, that satisfies such a condition, might be too hard to be discovered by an iterative scheme, especially if variations between the functions of the sequence $\{f_t\}_{t\in\mathbb{N}}$ are unknown or unpredictable. Thus assumptions on how the minimum changes in time or how the objective functions change in time will be required.

In general, without having access to the whole sequence, the best we can expect to find is that, under some condition on the sequence of objective functions, we can generate a sequence $\{x_t\}_{t\in\mathbb{N}}$ that tracks the time dependent minimum within a neighborhood. The next proposition will be illustrative of this fact. Before we state a lemma on recursive sequences that will be of use to break down proofs in a more agreeable manner

Lemma 2.4.1. Polyak (1987)
Let $\{u_t\}_{t\in\mathbb{N}}$ be a real valued sequence such that

$$u_{t+1} \leq q_0 u_t + q_1 u_{t-1} + \cdots + q_p u_{t-p} + \epsilon, \quad \epsilon > 0$$

for all t, where $p \in \mathbb{N}$ and $q_1, \ldots, q_p \in \mathbb{R}$ are such that the roots $r_0, r_1, \ldots, r_p$ of the associated characteristic equation

$$r^{p+1} - q_0 r^p - q_1 r^{p-1} - \cdots - q_p = 0$$

lie inside the unit circle. Then for fixed $k \in \mathbb{N}$

$$u_k \leq \frac{\epsilon}{1 - q_0 - q_1 - \cdots - q_p} + \left(\max_{n \in \{0,\ldots,p\}} \{r_n\} \right)^k a$$

where $a \in \mathbb{R}$.

This lemma shows that a recursive inequality assumption on the sequence $\{u_t\}_{t \in \mathbb{N}}$ is enough to show that it converges geometrically into the region

$$\{u \in \mathbb{R} | u < \epsilon/(1 - q_0 - q_1 - \cdots - q_p)|\}$$

with ratio $\max_{n \in \{0,\ldots,p\}} \{r_n\}$. We will use this lemma in the next proposition

Proposition 2.4.2. Polyak (1987)
Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t \in \mathbb{N}}$ be twice differentiable and satisfy assumptions 2.0.1 and 2.0.2.
Suppose also that $||x_t^* - x_{t+1}^*||_2 \leq a$. Then iteratively running the (time varying) gradient descent update $x_{t+1} = x_t - \alpha \nabla f_t(x_t)$ with $\alpha \leq 2/\mathrm{L}$ one obtains that

$$\limsup_{t \to \infty} ||x_t - x_t^*||_2 \leq \frac{a}{1 - q}, \qquad q = \max\{|1 - \alpha \ell|, |1 - \alpha \mathrm{L}|\} < 1.$$

This proposition shows how, subject to the assumption that the the minimum can only vary a certain amount, performing (time varying) gradient descent with a fixed step-size is indeed a sensible strategy to obtain an approximation that gets closer and closer to the minimum through time, albeit only in a neighborhood. Note how in this case we needed to use the lim sup since the sequence could oscillate in the neighborhood forever. Thus, by naively applying gradient descent with a time varying gradient, one manages to obtain convergence in a neighborhood of the time varying minimum, if the distances between adjacent time varying minima are uniformly bounded.
It is possible to weaken the assumption on the bounds of the increase in the minima $||x_t^* - x_{t+1}^*||$ to obtain a similar proposition

Proposition 2.4.3. Popkov (2005)
Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t \in \mathbb{N}}$ be twice differentiable and satisfy assumptions 2.0.1 and 2.0.2. Suppose also that $||\nabla f_{t+1}(x) - \nabla f_t(x)|| \leq h$ for all t, where $h > 0$. Then iteratively running the (time varying) gradient descent update $x_{t+1} = x_t - \alpha \nabla f_t(x_t)$ with $\alpha \leq 2/\mathrm{L}$ one obtains that

$$\limsup_{t \to \infty} ||x_t - x_t^*||_2 \leq \frac{h}{\ell(1 - q)}, \qquad q = \max\{|1 - \alpha \ell|, |1 - \alpha \mathrm{L}|\} < 1.$$

The assumption $||\nabla f_{t+1}(x) - \nabla f_t(x)|| \leq h$ in fact implies, for a strongly convex, function that $||x_t^* - x_{t+1}^*||$ will be uniformly bounded.

2.4.1 A Prediction Correction Algorithms

Up until now we have considered a modification of the gradient descent scheme by using the gradient of the objective function at time t 2.4. This kind of algorithm falls under the class of running algorithms (or correction only/catching up) because it does not attempt to predict where the next optimizer will be but it bases its update only on the gradient at time t.

Instead of utilizing a running algorithm, like the naive modification of gradient descent 2.4, there are other options to tackle the time varying case that result in more sophisticated algorithms. Specifically, there exist algorithms that attempt to infer how the sequence of optimizers are changing in time and only then do they apply a corrective step. Schemes that follow this principle are called prediction-correction algorithms. One of the most recently theorized ones is the Approximate Gradient Tracking (AGT) Simonetto et al. (2016). We present it here since just how score driven models were inspired by schemes involving second order expansions of the log observation density Creal et al. (2013) we will later argue that other observation driven models may retain desirable properties by taking inspiration from sophisticated optimization algorithms.

Approximate Gradient Tracking

```
Require: initial variable $x_0$, Initial objective function $f_{t_0}(x)$, no. of correction steps $\tau$.
  for $k = 0, 1, 2, \dots$ do
          $x_{k+1|k} = x_k - h\left[\nabla_{xx} f_{t_k}(x_k)\right]^{-1} \tilde{\nabla}_{xt} f_{t_k}(x_k)$        (2.6)
      Acquire the updated function $f_{t_{k+1}}(x)$.
      Initialize the sequence of corrected variables $\hat{x}_{k+1}^0 = x_{k+1|k}$
      for $q = 0 : \tau - 1 \dots$ do
              $\hat{x}_{k+1}^{q+1} = \hat{x}_{k+1}^q - \alpha \nabla_x f_{t_{k+1}}(\hat{x}_{k+1}^q)$        (2.7)
      end for
      Set the corrected variable $x_{k+1} = \hat{x}_{k+1}^\tau$
  end for
```

The AGT is a prediction-correction scheme: first there is a prediction step given by 2.6

then a correction step given by 2.7. The aim of the prediction step is to keep the gradient approximately constant while the optimization problem is changing in time. A brief explanation on how the prediction step is recovered may be given as so: the evolution of the gradient viewed as a function of both t and x is approximately given by

$$\nabla_x f_{t+\delta t}(x+\delta_x) \approx \nabla_x f_t(x) + \nabla_{xx} f_t(x)\delta_x + \nabla_{tx} f_t(x)\delta_t$$

Then by imposing the equality $\nabla_x f_{t+\delta t}(x+\delta_x) = \nabla_x f_t(x)$ we are left with

$$\frac{\delta_x}{\delta_t} = \frac{\nabla_{tx} f_t(x)}{\nabla_{xx} f_t(x)}$$

that is tantamount to the continuous dynamical system

$$\dot{x} = \frac{\nabla_{tx} f_t(x)}{\nabla_{xx} f_t(x)}. \tag{2.8}$$

This motivates the discrete version

$$x_{k+1|k} = x_k - h\left[\nabla_{xx} f_{t_k}(x_k)\right]^{-1} \nabla_{xt} f_{t_k}(x_k) \tag{2.9}$$

where $h := t_k - t_{k-1}$ and in the prediction step, 2.6, $\nabla_{xt} f_{t_k}(x_k)$ is approximated by

$$\tilde{\nabla}_{xt} f_{t_k}(x_k) := \frac{1}{h}\left(\nabla_{xt} f_{t_k}(x_k) - \nabla_{xt} f_{t_{k-1}}(x_k)\right) \tag{2.10}$$

since in applied setting we rarely have access to the variation of the objective function over time. Thus the prediction step given by the variable $x_{k+1|k}$ approximately maintains the direction and the magnitude of the gradient at the previous time, i.e., $\nabla_x f_{t+1}(x_{k+1|k}))$ should be close to $\nabla_x f_t(x_k)$.

On the other hand, the correction step 2.7 is based on the gradient descent method to correct the predicted decision variable $x_{k+1|k}$. This procedure modifies the predicted variable $x_{k+1|k}$ towards the optimal argument of the objective function at time t_{k+1}, $\alpha > 0$ is the step-size. Notice that the correction step, given by gradient descent, requires the updated objective function $f_{t_{k+1}}(x)$.

To prove convergence of the AGT scheme, some technical conditions are required.

Assumption 2.4.4.

The function $f_t(x)$ is twice differentiable and ℓ-strongly convex in $x \in \mathbb{R}^n$ and uniformly in t, that is

$$\ell I \preceq \nabla_{xx} f_t(x), \qquad \forall x \in \mathbb{R}^n, t.$$

This assumption does not only guarantee an unique minimum for each objective function,

but also ensures that the hessian of the objective function $f_t(x)$ is invertible.
This is analogous to Assumption 2.0.1 in the time varying objective function case; as a matter of fact it is the same as imposing Assumption 2.0.1 for all t, if the objective function is twice differentiable.

Assumption 2.4.5.
the function $f_t(x)$ has bounded second and third order derivatives with respect to $x \in \mathbb{R}^n$ and t, i.e.,

$$\begin{aligned}&||\nabla_{xx} f_t(x)|| \leq L, \ ||\nabla_{tx} f_t(x)|| \leq C_0, \ ||\nabla_{xxx} f_t(x)|| \leq C_1, \\ &||\nabla_{xtx} f_t(x)|| \leq C_2, \ ||\nabla_{ttx} f_t(x)|| \leq C_3\end{aligned}$$

This assumption ensures three facts: the Lipschitz continuity of the gradient at each time t, that the third derivative $\nabla_{xxx} f_t(x)$ is bounded above (this is typically required when dealing with the convergence of a Newton type algorithm) and the boundedness of the time variations of both the gradient and the Hessian.
To start the convergence analysis it will be appropriate to give a definition and a preliminary lemma

Definition 2.4.6.
The approximation error of the first-order forward Euler integral in 2.9 with respect to the dynamics in 2.8 is given by

$$\Delta_k := x_{k+1|k} - x(t_{k+1})$$

where $x(t_{k+1})$ is the exact prediction obtained by integrating the continuous dynamics in 2.8 from the initial condition x_k and $x_{k+1|k}$ is that of 2.9 with the correct mixed gradients.

An upper bound for $||\Delta_k||$ will be central in proving convergence of the algorithm since it will constrain the error coming from the prediction step.

Lemma 2.4.7. Simonetto et al. (2016)
Under assumptions 2.4.4, 2.4.5, the norm of the approximation error $||\Delta_k||$ is bounded from above by

$$||\Delta_k|| \leq \frac{h^2}{2}\left[\frac{C_0^2 C_1}{\ell^3} + \frac{2C_0C_2}{\ell^2} + \frac{C_3}{\ell}\right] = O(h^2)$$

We have that the norm of the approximation error $||\Delta_k||$ is bounded above by a constant that is of the order of $O(h^2)$.
Now we may prove that AGT converges exponentially to a neighborhood of the true value at time k.

Proposition 2.4.8. Simonetto et al. (2016)
Let assumptions 2.4.4 and 2.4.5 hold and define the constants

$$\begin{aligned}&\rho := \max\{|1-\alpha\ell|, |1-\alpha L|\}, \quad \sigma := 1 + h(C_0C_1/\ell^2 + C_2/\ell),\\ &\Gamma := h^2/2\left[C_0^2C_1/\ell^3 + 2C_0C_2/\ell^2 + 2C_3/\ell\right].\end{aligned}$$

Then by choosing the step-size $0 < \alpha < 2/L$, which implies $\rho < 1$, we have that, for any sampling period h, iteratively running AGT results in

$$||x_{k+1} - x^*(t_{k+1})|| \leq \rho^{\tau(k+1)}||x_k - x^*(t_k)|| + \rho^\tau \left[h\frac{2C_0}{\ell} + \Gamma\right]\left[\frac{1-\rho^{\tau(k+1)}}{1-\rho^\tau}\right],$$

and if we choose the sampling period h such that $\rho^\tau\sigma < 1$ then we obtain the bound

$$||x_k - x^*(t_k)|| \leq (\rho^\tau\sigma)^k||x_0 - x^*(t_0)|| + \rho^\tau\Gamma\left[\frac{1-(\rho^\tau\sigma)^k}{1-\rho^\tau\sigma}\right].$$

Thus the AGT prediction-correction method can always attain an error bound of the order of $O(h^2)$ if the sampling period is small enough. Moreover, in Simonetto et al. (2016), a series of numerical experiments are made which show that it is indeed possible to obtain bounds on the neighborhood of convergence of the order of $O(h^4)$ in a majority of cases.

Having viewed the AGT it is tempting to imagine a different prediction step that, instead of trying to keep the evolution of the gradient constant, attempts to keep the value of the function $f_t(x)$ constant through time. In fact an approximation of the dynamics through time of the type

$$f_{t+\delta t}(x + \delta_x) \approx f_t(x) + \nabla_x f_t(x)\delta_x + \nabla_t f_t(x)\delta_t$$

results in an approximation step given by

$$x_{k+1|k} = x_k - h\left[\nabla_x f_{t_k}(x_k)\right]^{-1}\nabla_t f_{t_k}(x_k).$$

But, while in the case of a strongly convex function with Lipschitz continuous gradient, we know that the second derivative is bounded both from above and below $0 < I\ell \leq \nabla_{xx}f_{t_k}(x_k) \leq IL$, the first derivative $\nabla_x f_{t_k}(x_k)$ is generally unbounded and can even be equal to zero. In addition imposing bounds on the first derivative is quite restrictive in that it excludes even a quadratic objective function from the analysis.

2.4.2 The Dynamical System Assumption

An original assumption that will be explored in this thesis is concerned with the evolution of the sequence of minima through time, that is, what if one knew that the sequence of minima was given by a dynamical system

$$x^*_{t+1} = \phi(x^*_t)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a known function, on which we will impose regularity conditions. The problem then becomes: with this ulterior assumption can we recover an optimization scheme that converges to the time varying minimum. Note how the problem has not been trivialized; because even knowing the exact evolution of the minimum one does not know the starting value x^*_1 of the dynamical system. Thus the entire sequence of time varying minima can't be recovered a priori, a situation that resembles the invertibility problem in econometric models (we can imagine the dynamical system as a model and the scheme as a filter, this parallel will be discussed more in chapter 4).

Although we choose to analyze this problem with an eye to applications in the theory of econometric model building, the problem itself, i.e., finding a scheme that converges given the dynamics of the minima x^*_{t+1} but without the starting value is an interesting mathematical investigation of its own. This section will be dedicated to exploring time varying gradient descent in this setting.

First we will restrict ourselves to the one dimensional case with simple dynamical systems of the kind

$$x^*_{t+1} = \omega + \theta x^*_t \tag{2.11}$$

where $\omega \in \mathbb{R}, \theta \in \mathbb{R}^+$ and $x^*_t \in \mathbb{R}$ for all t.

Given that we know how the minimum varies in time we expect to be able to modify the optimization scheme in a way that takes advantage of this extra information. In particular, a natural modification of gradient descent in this setting is

$$x_{t+1} = \omega + \theta x_t + \alpha \nabla f_{t+1}(\omega + \theta x_t) \tag{2.12}$$

where $\alpha \in \mathbb{R}$ is the step hyper-parameter.

In this way at each step of the algorithm we update x_{t+1} in the direction of the time varying minima by applying the same dynamics, the minima possesses, to x_t; resulting in $\omega + \theta x_t$. Then we direct it closer to the minimum x^*_{t+1}, by utilizing the direction given to us by the gradient at that point $\nabla f_{t+1}(\omega + \theta x_t)$.

In an applied setting one may wish to track x^*_{t+1} given the gradient at the previous time ∇f_t only. This restriction appears in prediction problems where we posses knowledge only up to the previous time. Thus we will also be interested in the scheme

$$x_{t+1} = \omega + \theta x_t + \alpha \nabla f_t(x_t) \tag{2.13}$$

where we compute the gradient at the previous time with respect to x_{t+1} (that tells us the direction in which we have to go to reach x^*_t). Given the simple dynamics of x^*_{t+1} we will see that this is enough to achieve convergence.

For the scheme given by 2.12 we achieve the following result

Proposition 2.4.9.
Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t\in\mathbb{N}}$ be twice differentiable and satisfy assumptions 2.0.1 and 2.0.2 and let x^*_t evolve according to equation 2.11 where $\theta < (\ell + L)/(L - \ell)$. Then choosing $(\theta - 1)/\theta\ell < \alpha < (\theta + 1)/\theta L$ the scheme given by 2.12 will be such that

$$|x_{t+1} - x^*_{t+1}| \leq q|x_t - x^*_t|$$

where $0 < q < 1$.

We see that, to obtain convergence some bounds must be imposed on θ, as expected, since x^*_{t+1} distances itself from previous values more than linearly, while the correction provided by the gradient is linear in nature. This statement in particular allows θ, depending on ℓ and L, to be greater than one and thus the sequence $\{x^*_{t+1}\}_{t\in\mathbb{N}}$ to not converge. This is a most interesting case since if $\{x^*_{t+1}\}_{t\in\mathbb{N}}$ were convergent regardless of its starting value x^*_0 then recovering a scheme such that $|x_{t+1} - x^*_{t+1}|$ were to converge would be trivial.

Surprisingly, an analogous statement can be made for the scheme in 2.13.

Proposition 2.4.10.
Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t\in\mathbb{N}}$ be twice differentiable and satisfy assumptions 2.0.1 and 2.0.2 and let x^*_t evolve according to equation 2.11 where $\theta < (\ell + L)/(L - \ell)$. Then choosing $(\theta - 1)/\ell < \alpha < (\theta + 1)/L$ the scheme given by 2.13 will be such that

$$|x_{t+1} - x^*_{t+1}| \leq q|x_t - x^*_t|$$

where $0 < q < 1$.

This fact tells us that what matters in the end in order to achieve convergence is the direction given to us by the gradient. In fact the direction of the gradient will always point towards the minimum at time $t + 1$, thanks to the simple dynamics of x^*_{t+1}, that imply that whichever side x_t lies on, with respect to x^*_t, so will $\theta x_t + \omega$, with respect to x^*_{t+1}.

Now, with an eye to our original dynamical system assumption, we analyze the more general setting when x^*_{t+1} evolves according to a dynamical system of the kind

$$x^*_{t+1} = \phi(x^*_t) \tag{2.14}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a differentiable function and $x^*_t \in \mathbb{R}$. With this recurrent equation for x^*_{t+1} we will choose, following the intuition of the simpler cases discussed before, the gradient scheme as

$$x_{t+1} = \phi(x_t) + \nabla f_t(x_t) \tag{2.15}$$

or, as before,

$$x_{t+1} = \phi(x_t) + \nabla f_{t+1}(\phi(x_t)) \tag{2.16}$$

if the gradient at time $t+1$ is assumed known at time t. Although, in the remainder of this section we will only analyze the scheme given by 2.15, reason being that analogous proofs following the exact same ideas can be given for the scheme in 2.16.

Given the results in 2.4.9 and 2.4.10 we know that we must impose some conditions on ϕ if we want to obtain convergence of the scheme 2.15. Based on some strong conditions on the derivative of ϕ we are able to obtain a convergence result.

Proposition 2.4.11.

Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t\in\mathbb{N}}$ be twice differentiable and satisfy assumption 2.0.1 and 2.0.2, let x^*_t evolve according to equation 2.14 and assume that $c - \epsilon < \phi'(x) < c + \epsilon$ for all $x \in \mathbb{R}$ where $0 < \epsilon < 1$ and $c < (\ell + L)(-\epsilon + 1)/(L - \ell)$. Then choosing $(c + \epsilon - 1)/\ell < \alpha < (c - \epsilon + 1)/L$ the scheme given by 2.15 will be such that

$$|x_{t+1} - x^*_{t+1}| \leq q|x_t - x^*_t|$$

where $0 < q < 1$.

In particular notice that the bounds imposed on the derivative of ϕ make the function ϕ itself Lipschitz, allowing us to bound the distance between two outputs of the function with the inputs as utilized in the proof.

Finally we wish to tackle the problem when the recursive equation 2.14 is multidimensional, that is when

$$x^*_{t+1} = \phi(x^*_t) \tag{2.17}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a differentiable function and $x^*_t \in \mathbb{R}^d$.

To generalize the same arguments as before to the d-th dimension we will need the gradient of ϕ to be symmetric, an ulterior restrictive assumption.

Proposition 2.4.12.

Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t\in\mathbb{N}}$ be twice differentiable and satisfy assumption 2.0.1 and 2.0.2, let x_t^* evolve according to equation 2.17 and assume that $\nabla\phi(x)$ is symmetric and such that $I(c-\epsilon) < \nabla\phi(x) < I(c+\epsilon)$ for all $x \in \mathbb{R}$ where $0 < \epsilon < 1$ and $c < (\ell+L)(-\epsilon+1)/(L-\ell)$. Then choosing $(c+\epsilon-1)/\ell < \alpha < (c-\epsilon+1)/L$ the scheme given by 2.15 will be such that

$$\|x_{t+1} - x_{t+1}^*\| \leq q\|x_t - x_t^*\|$$

where $0 < q < 1$.

To achieve this result a bound on the eigenvalues of the sum of two Hermitian matrices, knowing the eigenvalues of both summands, was needed. As elementary as this problem may seem to state the complete resolution was only given in 2001 by Knutson and Tao (2001). Interestingly, the corresponding problem with generic matrices is still unsolved. Notice also that the matrix inequality is not well defined when a matrix is non symmetric. Of course there exist other gradient based schemes that could work better in this setting, in particular schemes that converge faster to the minimum like Nestervo's accelerated gradient descent Nesterov (1983) or Polyak's momentum Polyak (1987). The results in this section are nothing but a first dip in the theory and are meant as not much more than examples.

Before ending this section we consider a final case: assume the sequence of minima obeys the dynamic equation

$$x_{t+1}^* = \omega + \theta_1 x_t^* + \theta_2 x_{t-1}^* + \cdots + \theta_{p+1} x_{t-p}^* \tag{2.18}$$

with coefficients $\omega, \theta_1, \ldots, \theta_{p+1} \in \mathbb{R}$ and we take $x_t^* \in R$. Then we can easily obtain convergence of the scheme given by

$$x_{t+1} = \omega + \theta_1 x_t - \alpha_1 \nabla f_t(x_t) + \theta_2 x_{t-1} - \alpha_2 \nabla f_{t-1}(x_{t-1}) + \cdots + \theta_{p+1} x_{t-p} - \alpha_p \nabla f_{t-p}(x_{t-p}) \tag{2.19}$$

assuming that the recurrent sequence in 2.18 has all roots of its associated characteristic equation inside the unit circle. In fact we state the following result

Proposition 2.4.13.

Let every element $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t\}_{t\in\mathbb{N}}$ be twice differentiable and satisfy assumptions 2.0.1 and 2.0.2, let x_t^* evolve according to equation 2.18 that has all roots of its characteristic equation inside the unit circle. Then there exist $\alpha_1, \ldots, \alpha_p$ such that the scheme given by 2.19 converges, i.e.,

$$\|x_{t+1} - x_{t+1}^*\| \to 0.$$

Although in this proposition we only state that the α_i exist for small values of p specific bounds where the α_i must lie in can be found analytically, and for large values of p they can be found numerically. The assumption on the roots of the characteristic equation of 2.18 implies that the sequence of the x_t^* converges to zero. A case that is not particularly interesting, but the way the proof is formulated clearly shows that the assumption can be weakened on a case by case basis. In general we can obtain convergence of schemes of the form 2.19 under the assumption that equation 2.18 does not have all roots of the associated characteristic equation inside the unit circle.

2.4.3 Adding a Stochastic Element to the Time Varying Objective Functions

In this section we briefly consider a further generalization of the time varying optimization problem that consists in considering a sequence $\{f_t(x,\epsilon_t)\}_{t\in\mathbb{N}}$ of time varying objective functions that also depend on a random variable ϵ_t at each time t. Because of the random component we will be interested in the expected update given by the gradient descent. Supposing that the time varying minimum evolves according to some discrete stochastic process, it is possible to obtain propositions like the following

Proposition 2.4.14.
Let every $f_t : \mathbb{R}^d \to \mathbb{R}$ of the sequence $\{f_t(x,\epsilon_t)\}_{t\in\mathbb{N}}$ satisfy assumptions 2.0.1 and 2.0.2 and assume the sequence of minima $\{x_t^*\}_{t\in\mathbb{N}}$ evolves according to a stationary AR(1) process, i. e. $x_{t+1}^* = \omega + \beta x_t^* + \epsilon_t$, where $\omega \in \mathbb{R}$, $-1 < \beta < 1$ and ϵ_t is a IID white noise process with zero mean and constant variance $\sigma_{\epsilon_t}^2$. Then iteratively running k times the (time varying) gradient descent update $x_{t+1} = \omega + \beta x_t - \alpha \nabla f_t(x_t)$, with $(-1+\beta)/\ell \le \alpha \le (1+\beta)/\mathrm{L}$, one obtains that

$$\lim_{t\to\infty} E_{\epsilon_t}[||x_{t+1}^* - x_{t+1}||] \le \frac{\epsilon}{1-q}, \qquad q = \max\{|\beta - \alpha\ell|, |\beta - \alpha\mathrm{L}|\} < 1$$

where $\epsilon = E_{\epsilon_t}[||\epsilon_t||]$.

Here we notice that, even if the points where the sequence of objective functions reach the minima vary in a stochastic fashion, we can still adjust the gradient descent update to obtain a scheme that in expectation lies within a neighborhood of the time varying minimum for large t. Moreover one can give explicit bounds for this expected neighborhood. As before we could obtain a more general result for non stationary auto-regressive

processes following a similar proof as in section 2.4.2. In general all results of the previous section can be extended to this simple stochastic setting.
The next proposition deals with a very specific sequence of objective functions

Proposition 2.4.15.
Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ satisfy assumption 2.0.1 and 2.0.2.
Construct the sequence $\{f_t\}_{t\in\mathbb{N}}$ given by $f_{t+1}(x) = f_t(x + \epsilon_t)$ where ϵ_t is white noise and has variance $\sigma_t^2 < M$ for all t where $M \in \mathbb{R}$.
Then running the iteration $x_{t+1} = x_t - \alpha\nabla f_{t+1}(x_t)$, from an initial starting point x_0, for k times, with a fixed step size $\alpha \leq 1/L$, we obtain that

$$\lim_{t\to\infty} E[f_t(x_t) - f_t(x_t^*)] \leq \frac{M}{\ell\alpha}. \tag{1.1}$$

This result tells us that for a sequence of objective functions, that is shifted in space like a random walk, the time varying gradient descent scheme converges in expectation to a neighborhood of the true value.

Chapter 3

Observation Driven Models

Since the categorization given by Cox et al. (1981) observation driven models have multiplied profusely and are increasingly being adopted to model time varying parameters. These models are widely applied in various fields ranging from economics, see Pindyck and Rubinfeld (1998), environmental study, see Bhaskaran et al. (2013), epidemiology and public health study, see Zeger (1988), Ferland et al. (2006) and finance where the celebrated GARCH model is an industry standard for the modeling of volatility Bollerslev (1986).

For observation driven models the filtering (or dynamic) equation is specified as a function of past observations as well as contemporaneous and lagged exogenous variables. Thus, although the parameters are stochastic, they are perfectly predictable given past information. With an observation driven specification the likelihood is readily available, hence the parameter estimation is relatively simple, and prediction is straightforward.

There are many definitions of observation driven models in the literature with varying degree of associated technicalities. We will broadly follow the definition given in Douc et al. (2013) that highlights the conditional nature of the model specification. To introduce the models we utilize the notation $x_{l:m} := \{x_l, \dots, x_m\}$.

A stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$ valued in a measurable space of choice $(\mathtt{Y}, \mathcal{Y})$ is said to be an observation driven model of order (p, q) (ODM(p,q)) if there exists a process $\{\tilde{\lambda}_t\}_{t\in\mathbb{Z}}$ taking values in a measurable space $(\tilde{\Lambda}, \tilde{\Lambda})$ such that for all $t \in \mathbb{Z}$

$$\begin{aligned} &Y_t|\mathcal{F}_{t-1} \sim \tilde{p}(y_t|\tilde{\lambda}_t, \boldsymbol{\theta}), \\ &\tilde{\lambda}_{t+1} = \phi(\tilde{\lambda}_{t-p+1:t}, Y_{t-q+1:t}, \boldsymbol{\theta}), \end{aligned} \tag{3.1}$$

where $\tilde{p}(\cdot|\tilde{\lambda}_t, \boldsymbol{\theta})$ is a parametric conditional density, $\mathcal{F}_t := \sigma(Y_l; l \leq t; l \in \mathbb{Z})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ is a vector of static unknown parameters and ϕ is a measurable function, also called the updating function, linking the new $\tilde{\lambda}_{t+1}$ to the current and past observations. Specifically it is assumed that $\tilde{\lambda}_t$ is a measurable function of $Y^{t-1} := \{Y_{t-1}, Y_{t-2}, \dots, Y_1, \dots\}$

and $\boldsymbol{\theta}$, for some initial value $\bar{\lambda}_1 := (\bar{\lambda}_{11}, \bar{\lambda}_{12}, \dots, \bar{\lambda}_{1p})$ where each element belongs to $\tilde{\Lambda}$. The allure of observation driven models is that their likelihoods can be computed exactly, with computational complexity of the same order as the number of observations, making maximum likelihood estimation the privileged approach for statistical inference. The reason for the tilde above the probability density $\tilde{p}$ and the time varying parameter $\tilde{\lambda}_t$ is that we will often consider the misspecified case, that is when the time series $\{Y_t\}_{t\in\mathbb{Z}}$ actually has probability density $p(y_t|\lambda_t)$ different from $\tilde{p}(y_t|\tilde{\lambda}_t)$. Moreover even the case in which the recursive equation the time varying parameter follows 3.1 does not include the true functions that governs the evolution of the time varying parameter λ_t will be briefly considered.

For general observation driven models consistency and asymptotic normality of the model parameters $\boldsymbol{\theta}$ have been proven, even in the case of model misspecification, for a wide range of observation driven models as can be seen in Douc et al. (2013).

3.1 Score Driven Models

Score driven models are a subset of observation driven models, they have been introduced in the econometric literature in 2008 through the independent works of Creal et al. (2013) and Harvey and Chakravarty (2008).

The principal feature of these models is that the dynamics of the time varying parameter are driven by the score of the conditional distribution of the observations. Score-driven models are typically appreciated for the fact that they flexibly adapt themselves to the distribution of the innovations. Thus, choosing the distribution of the innovations as heavy tailed, one can easily produce models that have robustness properties. Moreover the analytic theory behind the maximum likelihood estimation of the parameters is quite developed, thanks to the form of observation driven models that allows to write down the likelihood function of the observations in a straightforward manner.

Several applications of score driven models have been carried out; Harvey and Luati (2014a) study the robustness of location type models, Creal et al. (2013) introduce a mixed measurement dynamic factor models, Bernardi and Bernardi (2018) expand the theory defining a two-sided skew and shape dynamic conditional score model, Blasques et al. (2016b) and Catania et al. (2018) apply the general theory to spatial models.

3.1.1 The Model Specification

A stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$, with associated filtration $\mathcal{F}_t := \sigma(Y_l; l \leq t; l \in \mathbb{Z})$, valued in $\mathbb{R}$ is said to be a score driven model of order (p, q) if there exists a process $\{\tilde{\lambda}_t\}_{t\in\mathbb{N}}$

taking values in $\tilde{\Lambda} \subseteq \mathbb{R}^n$ such that for all $t \in \mathbb{Z}$

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim \tilde{p}(y_t|\tilde{\lambda}_t, \boldsymbol{\theta}) \\ \tilde{\lambda}_{t+1} &= \gamma + \sum_{i=0}^{q} \alpha_i S_{t-i} s_{t-i} + \sum_{i=0}^{p} \beta_i \tilde{\lambda}_{t-i} \\ s_t &= \frac{\partial \log \tilde{p}(Y_t|\tilde{\lambda}_t, \theta)}{\partial \tilde{\lambda}_t} \end{aligned} \tag{3.2}$$

for some initial $\bar{\lambda}_1 \in \Lambda \subset \mathbb{R}^n$, where $p, q \in \mathbb{N}$, $\tilde{p}(y_t|\tilde{\lambda}_t, \boldsymbol{\theta})$ is a parametric conditional density , $\boldsymbol{\theta} \in \Theta \subseteq R^d$ is the vector of static parameters, and $S_t := S(\tilde{\lambda}_t, \boldsymbol{\theta})$ is a positive measurable scaling function that possibly depends on the filtered time-varying parameter $\tilde{\lambda}_t$ and the static parameter $\boldsymbol{\theta}$.

Notice that for probability densities that have a measurable score function

$$\tilde{\lambda} \to \partial \log \tilde{p}(Y_t|\tilde{\lambda}, \theta)/\partial \tilde{\lambda}$$

it is immediate to see that score driven models are a subset of observation driven models. The role of the score s_t in the updating equation for $\tilde{\lambda}_{t+1}$ has been a subject of interest in the literature. As illustrated in Creal et al. (2013), the score points in the steepest ascent direction for improving the model's local fit in terms of the likelihood at time t given the previous position of the parameter $\tilde{\lambda}_t$. We will formalize this intuition through the CEV property defined in section 4.2.

A more theoretically grounded motivation for the role of the score is given in Blasques et al. (2015), we will discuss it in detail in section 4.2 and build on its intuition to recover ulterior properties that score driven models posses, these will comprise most of the original results this thesis has to offer.

The selection of the scaling function S_t is also discussed in Creal et al. (2013) and will be object of our analysis. If chosen as the inverse of the information matrix

$$S_t := E_{Y_t|Y^{t-1}}[s_t s_t^T]^{-1} = -E_{Y_t}\left[\frac{\partial^2 \log \tilde{p}(Y_t|\tilde{\lambda}_t; \theta)}{\partial \tilde{\lambda}_t \partial \tilde{\lambda}_t^T}\right]$$

it allow one to obtain observation driven models such as the autoregressive conditional duration, Engle and Russell (1998), the autoregressive conditional intensity, Russell (2000), and the GARCH, Bollerslev (1986). Instead, when chosen as the identity matrix $S_t := I$, depending on the choice of distribution $\tilde{p}$ it is possible to recover models such as the autoregressive conditional multinomial model, Russell and Engle (2005), and the Beta-t-EGARCH model, Harvey and Chakravarty (2008). Thus the score driven specification is certainly wide and encompassing, capturing a great number of observation driven models already present in the literature.

We will argue that the scaling function S_t is not only useful in rendering the class of models more rich in number, but an accurate selection will guarantee desirable properties for the resulting score driven model.

3.1.2 The Beta-t-EGARCH Model and a Special Case

A special element of the class of score driven models is the generalized autoregressive conditional heteroskedasticity (1,1) (GARCH(1,1)) model that is usually specified by the equations

$$\begin{aligned} &Y_t = U_t\tilde{\sigma}_t \\ &\tilde{\sigma}^2_{t+1} = \omega + \alpha Y_t^2 + \beta\tilde{\sigma}^2_t \end{aligned} \tag{3.3}$$

where $(\omega, \alpha, \beta) := \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^3$ are the parameters and $\{U_t\}_{t\in\mathbb{N}}$ is a sequence of independent random variables with probability density $\tilde{p}_t$ that is Gaussian with zero mean and unitary variance.

The model is easily recovered in the score driven framework by noticing that

$$\frac{\partial \log \tilde{p}(Y_t|\tilde{\sigma}_t;\theta)}{\partial\tilde{\sigma}_t} = \frac{Y_t^2}{2\tilde{\sigma}_t^4} - \frac{1}{2\tilde{\sigma}_t^2}$$

thus choosing the positive scaling function $S_t(\tilde{\sigma}_t, \theta) = \tilde{\sigma}_t^4$ one recovers the GARCH(1,1) specification since

$$\tilde{\sigma}^2_{t+1} = \omega + \alpha S_t s_t + \beta\tilde{\sigma}^2_t = \omega + \frac{\alpha}{2}Y_t^2 + \tilde{\sigma}^2_t(\beta - \frac{\alpha}{2})$$

The primary use of the GARCH has been to capture the dynamics of volatility in financial markets, especially the volatility of financial returns. One of the purported shortcomings of the GARCH models is that, although financial returns are assumed distributed as a Gaussian, it is a stylized fact that financial returns have heavier tails, Bradley and Taqqu (2003). Besides, the GARCH model does not account for the leverage effect found in time series of financial returns (a stylized fact first noted by Black (1976)) that is, the model does not account for an asymmetric response in volatility to positive and negative shocks. Moreover, the GARCH model reacts excessively to one off spikes in returns as explained in Harvey (2013), in this sense the model lacks robustness.

It must be also said that despite the model in-adherence to some stylized facts it remains a standard in the financial industry and is able to, sometimes, outperform more complex models in comparative analysis Hansen and Lunde (2005b).

Although the GARCH model technically belongs to the class of score driven models it was specified before the score driven class was defined. A specification that takes full advantage of the flexibility of the score driven framework is the BETA-t-EGARCH model Harvey and Chakravarty (2008)

$$\begin{aligned} Y_t &= U_t \exp(\tilde{\sigma}_t^2) \\ \tilde{\sigma}_t^2 &= \omega + \alpha \frac{(\nu+1)Y_t^2}{\nu \exp(2\tilde{\sigma}_t^2) + Y_t^2} - \alpha + \beta \tilde{\sigma}_{t-1}^2 \end{aligned}$$

where $\{U_t\}_{t \in \mathbb{N}}$ is a sequence of independent random variables that have a generalized student-t probability density $\tilde{p}_t$ with zero mean, unitary location and ν degrees of freedom while $\gamma, \alpha, \beta \in \mathbb{R}$ are the unknown parameters. This model belongs to the score driven class since

$$\frac{\partial \log \tilde{p}(Y_t|\tilde{\sigma}_t;\theta)}{\partial \tilde{\sigma}_t} = \frac{(v+1)Y_t^2}{v \exp(2\tilde{\sigma}_t^2) + Y_t^2} - 1$$

where $\tilde{p}(\cdot|\tilde{\sigma}_t;\theta)$ is the probability density function of $Z_t \exp(\tilde{\sigma}_t^2)$. We explicitly notice that here the scaling function $S_t := 1$ plays no role.

The BETA-t-EGARCH model addresses some of the theoretical shortcomings of the GARCH model; it reacts less to one off spikes in returns, Harvey (2013), and it models Y_t with the t-student distribution that, depending on the degrees of freedom, is heavy tailed.

The framework that we will develop in section 4.2 will shed light on how S_t ought to be chosen to specify a model with additional properties as the choices made in the case of the GARCH and the BETA-t-EGARCH may appear ad hoc; made to recover the specification one desires. Instead, if the model is to posses ulterior desirable properties, the choice of the scaling function will be of importance in the model specification.

3.2 Maximum Likelihood Estimation and Invertibility

Although this thesis does not regard maximum likelihood estimation of parameters of score driven models, when introducing a statistical model it is best to also discuss if there exist valid methods to estimate the parameters of said model; otherwise the model risks lacking in practical usefulness. Besides, briefly presenting the most modern theory on maximum likelihood estimation regarding score driven models will introduce naturally the concept of model invertibility that will be object of a conjecture in 4.2.3.

As already stated, one of the main appeals of observation driven models is that, given a sequence of observations $y^t := \{y_t, y_{t-1}, \dots, y_1\}$, one can easily write down the log-likelihood function of the model as

$$\log \tilde{p}(y^t | y_0, \dots, y_{-p+1}, \tilde{\lambda}_0, \dots, \tilde{\lambda}_{-q+1}, \boldsymbol{\theta}) = \sum_{i=1}^{t} \log \tilde{p}(y_i | y_0, \dots, y_{-p+1}, \tilde{\lambda}_0, \dots, \tilde{\lambda}_{-q+1}, \boldsymbol{\theta}) \tag{3.4}$$

where the starting values $y_0, \dots, y_{-p+1}, \tilde{\lambda}_0, \dots, \tilde{\lambda}_{-q+1}$ have to be chosen to initialize the sequence given by 3.2. It is thus natural to estimate the vector of model parameters $\boldsymbol{\theta}$ through the method of maximum likelihood, i.e., to determine the value of $\hat{\boldsymbol{\theta}}$ as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \left\{ \sum_{i=1}^{t} \log \tilde{p}(y_i | y_0, \dots, y_{-p+1}, \tilde{\lambda}_0, \dots, \tilde{\lambda}_{-q+1}, \boldsymbol{\theta}) \right\}.$$

Historically, observation driven models have been sometimes introduced before proving the consistency and asymptotic normality of the maximum likelihood estimators of the parameters of said models, this was, for example, the case of the ARCH and GARCH models that were only tested empirically in their respective first papers Engle (1982), Bollerslev (1986). Only later proofs for consistency and asymptotic normality of the parameters of specific ARCH and GARCH models started to appear. First for the GARCH(1,1) case Lee and Hansen (1994a), Lumsdaine (1996b), and then for higher order observation driven models Strauman and Mikosch (2006).

However, consistency and asymptotic normality of the parameters, both in the correctly specified and misspecified case, are of tantamount importance to verify that the parameter estimates are meaningful. Although an asymptotic theory that governs all possible score driven specifications does not exist yet there are multiple results for large classes of them. We categorize these results in two sets: results based on Strauman and Mikosch (2006), and result based on lemma 1 of Jensen and Rahbek (2004). Strauman and Mikosch (2006) provide a general theory for handling estimation of non linear observation driven models assuming some restrictions on the parameter regions, crucially the asymptotic result they derive do not depend on the starting value $\tilde{\lambda}_0$, that is needed to write down the likelihood. This is attained by establishing the invertibility of the observation driven model under consideration, a concept we will soon define and discuss.

Lemma 1 in Jensen and Rahbek (2004) gives general conditions under which consistency and asymptotic normality hold, assuming that the true value of the unobserved time-varying parameter at time zero $\tilde{\lambda}_0$ is known. An assumption that is hardly ever true in applications.

Because invertibility will be connected to the ideas presented in section 4.2.2, in this thesis we will briefly present the state of the art theory based on Straumann and Mikosh, for an in depth exposition of the theory based on lemma 1 of Jensen and Rhabek one can check chapter 2 of Harvey (2013).

3.2.1 Maximum Likelihood for Score Driven Models

Due to the relative recent introduction of the score driven framework most of the asymptotic theory for score driven models, based on the ideas of Strauman and Mikosch (2006), can be found in three papers Blasques et al. (2016a), Blasques et al. (2014a), Blasques et al. (2014b).

Here we will briefly summarize only the most encompassing and recent results found in Blasques et al. (2014b), where the asymptotic theory of the maximum likelihood estimators of certain score-driven models is established. Guided by, but not limited to, first order models of the form

$$\begin{aligned} Y_t &= \tilde{\lambda}_t + U_t, & U_t &\sim \tilde{p}_U(u_t|\xi) \\ \tilde{\lambda}_{t+1} &= \omega + \alpha s_t + \beta \tilde{\lambda}_t, & s_t &:= \frac{\partial \log \tilde{p}_U(Y_t - \tilde{\lambda}_t|\xi)}{\partial \tilde{\lambda}_t} \end{aligned} \tag{3.5}$$

where Y_t takes values in a measurable space $(\mathtt{Y}, \mathcal{Y})$, $\tilde{\lambda}_t$ takes values in a measurable space $(\tilde{\Lambda}, \tilde{\Lambda})$, $\{U_t\}_{t\in\mathbb{Z}}$ is a sequence of random variables defined on a probability space $\{U, \mathcal{U}, \tilde{P}\}$ with probability density $\tilde{p}_U(u_t, \xi)$ for all t. The probability density $\tilde{p}_U(u_t, \xi)$ depends on an unknown static parameter vector ξ that for expository convenience is assumed to belong to a subset $\Xi \subseteq \mathbb{R}$ but can be chosen to belong to a set of arbitrary dimension and the results would still follow. The static parameter vector $\boldsymbol{\theta}$ in this setting is thus $\boldsymbol{\theta} := (\omega, \alpha, \beta, \xi) \in \Theta \subseteq \mathbb{R}^4$.

The general framework used to state the propositions is the following: consider a stochastic process $\{Y_t\}_{t\in N}$ given by

$$Y_t = g(\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1), U_t), \qquad U_t \sim \tilde{p}(u_t|\xi)$$

where $g : \tilde{\Lambda} \times U \to \mathtt{Y}$ is a link function that is strictly increasing in its second argument, $\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)$ is the time varying parameter which belongs to $\tilde{\Lambda}$ that is assumed to be a convex set in $\mathbb{R}$, $\{U_t\}_{t\in\mathbb{N}}$ is an exogenous i.i.d. sequence of random variables for every parameter vector $\xi \in \Xi \subseteq \mathbb{R}$.

The time varying parameter updating scheme is given by

$$\tilde{\lambda}_{t+1}(\boldsymbol{\theta}, \bar{\lambda}_1) = \omega + \alpha s(\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1), Y_t, \xi) + \beta \tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)$$

for $t > 1$, and initialized at $\tilde{\lambda}_1(\boldsymbol{\theta}, \bar{\lambda}_1) = \bar{\lambda}_1$ for a non-random $\bar{\lambda}_1 \in \tilde{\Lambda}$. The function $s : \tilde{\Lambda} \times \mathtt{Y} \times \Xi \to \tilde{\Lambda}$ is the scaled score of the conditional density of Y_t given $\tilde{\lambda}_t$. When there will be no possibility of confusion we will suppress the dependence of $\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)$ on its arguments and write $\tilde{\lambda}_t$ instead. We also define $\tilde{p}_Y(y_t|\tilde{\lambda}_t, \xi)$ as the conditional density of Y_t given $\tilde{\lambda}_t$

$$\tilde{p}_Y(y_t|\tilde{\lambda}_t, \xi) = \tilde{p}_U(\bar{g}(\tilde{\lambda}_t, Y_t)|\xi)\left(\frac{\partial \bar{g}(\tilde{\lambda}_t, Y_t)}{\partial Y_t}\right)$$

where $\bar{g}_t := \bar{g}(\tilde{\lambda}_t, Y_t) := g^{-1}(\tilde{\lambda}_t, Y_t)$ is the inverse of $g(\tilde{\lambda}_t, U_t)$ with respect to U_t.
The scaled score is thus given by $s(\tilde{\lambda}_t, Y_t, \xi) = S_t(\tilde{\lambda}_t, Y_t)s_t(\tilde{\lambda}_t, Y_t, \xi)$ where

$$s_t(\tilde{\lambda}_t, Y_t, \xi) = \left[\frac{\partial \bar{p}_t}{\partial \tilde{\lambda}_t} + \frac{\partial \log \bar{g}_t'}{\partial \tilde{\lambda}_t}\right]$$

with $\bar{p}_t := \bar{p}(\tilde{\lambda}_t, Y_t, \xi) = \log \tilde{p}_U(\bar{g}(\tilde{\lambda}_t, Y_t)|\xi)$, $\bar{g}_t' = \partial \bar{g}(\tilde{\lambda}_t, Y_t)/\partial Y_t$ and $S_t : \tilde{\Lambda} \times \Xi \to \tilde{\Lambda}$ is the positive scaling function.
Given a sequence of observations $\{y_i\}_{i=1}^T$ The average log-likelihood function $\mathcal{L}_T$ for a time series of the type in 3.5 is found in closed form as

$$\tilde{\mathcal{L}}_T(\boldsymbol{\theta}, \bar{\lambda}_1) := \frac{1}{T}\sum_{t=1}^{T}\Big(\log \tilde{p}_U(\bar{g}(\tilde{\lambda}_t, y_t)|\xi) + \log \partial \bar{g}(\tilde{\lambda}_t, y_t)/\partial y_t\Big)$$

and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1)$ is given by

$$\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1) := \arg\max_{\boldsymbol{\theta} \in \Theta} \tilde{\mathcal{L}}_T(\boldsymbol{\theta}, \bar{\lambda}_1)$$

where we have explicitly remarked the dependence on the initial condition $\bar{\lambda}_1$. In the case of correct model specification, that is when the assumed probability distribution $\tilde{p}$ is equal to the true distribution p that governs the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$, we will drop the tilde, i.e., $\mathcal{L}_T$ will denote the log likelihood of the correctly specified model.
To derive the maximum likelihood properties Blasques et al. (2014b) analyze the stochastic behavior of the filtered time-varying parameter over different $\boldsymbol{\theta} \in \Theta$. The upcoming proposition is the main argument used in establishing the exponentially fast almost sure convergence of the score-driven filtered sequence $\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)$ uniformly over the parameter space Θ. To state the proposition concisely, we define $\dot{s}_{Y_t}(\tilde{\lambda}, \xi) := \partial s(\tilde{\lambda}, Y_t, \xi)/\partial \tilde{\lambda}$ and

$$\bar{\rho}_t^k(\boldsymbol{\theta}) := \sup_{\tilde{\lambda} \in \tilde{\Lambda}} |\beta + \alpha \dot{s}_{Y_t}(\tilde{\lambda}, \xi)|^k$$

Proposition 3.2.1. Blasques et al. (2014b)
Let $\Theta \subseteq \mathbb{R}^4$ be compact, $s \in C^{(1,0,0)}(\tilde{\Lambda} \times \mathtt{Y} \times \Xi)$ and let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary and

ergodic stochastic process. Assume there exist $\bar{\lambda}_1 \in \tilde{\Lambda}$ such that

$$(i)\;\; E\left[\log \sup_{\tilde{\lambda}\in\tilde{\Lambda}} |s(\bar{\lambda}_1, Y_t, \xi)|\right]^{+} < \infty$$

$$(ii)\;\; E\left[\log \sup_{\boldsymbol{\theta}\in\Theta} \bar{\rho}_1^1(\boldsymbol{\theta})\right] < 0$$

where $x^+ := \max(0, x)$.
Then uniformly on Θ the sequence $\{\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)\}_{t\in\mathbb{N}}$ converges exponentially almost surely to a unique stationary and ergodic sequence $\{\tilde{\lambda}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ as $t \to \infty$, i.e.,

$$c^t \|\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1) - \tilde{\lambda}_t(\boldsymbol{\theta})\| \overset{a.s.}{\to} 0$$

for some $c > 1$.

This result is crucial in the proof of uniform convergence of the estimators through the application of the ergodic theorem of Rao for sequences in separable Banach space Rao (1962). This allow one to obtain consistency and asymptotic normality under weaker differentiability conditions like the ones found in Strauman and Mikosch (2006).
We now state the simplified assumptions to obtain consistency

Assumption 3.2.2.
Let $(\Theta, \mathcal{B}(\Theta))$ be a measurable space with Θ compact.

Assumption 3.2.3.
Let $s \in C^{(2,0,2)}(\tilde{\Lambda} \times \mathtt{Y} \times \Xi)$.

Assumption 3.2.4.
There exists $\Theta^* \in \mathbb{R}^4$, $m_{\tilde{\lambda}} > 0$ and $\delta > 0$ such that for any $\bar{\lambda}_1 \in \tilde{\Lambda}$

$$(i)\;\; ||s(\bar{\lambda}_1, Y_t, \cdot)||^{\Theta^*}_{m_{\tilde{\lambda}}+\delta} < \infty$$

$$(ii)\;\; \sup_{(\tilde{\lambda}, y, \boldsymbol{\theta})\in\tilde{\Lambda}\times\mathtt{Y}\times\Theta^*} |\beta + \alpha \partial s(\tilde{\lambda}, y, \xi)/\partial\tilde{\lambda}| < 1$$

where $||s(\bar{\lambda}_1, Y_t, \cdot)||^{\Theta^*}_{m_{\tilde{\lambda}}+\delta} := \left(E[\sup_{\boldsymbol{\theta}\in\Theta^*} |s(\bar{\lambda}_1, Y_t, \boldsymbol{\theta})|^{m_{\tilde{\lambda}}+\delta}]\right)^{1/(m_{\tilde{\lambda}}+\delta)}$.

Assumptions 3.2.2 is a standard assumptions in the maximum likelihood literature where the parameter space is usually required to be compact. Assumption 3.2.3 guarantees a well behaved function s, eliminating the possibility for degenerate cases, while assumption 3.2.4 ensures the convergence of the sequence $\{\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)\}_{t\in\mathbb{N}}$ to a stationary and ergodic one.
For the next assumption it will be convenient to define a piece of notation: for any differentiable function f that depends on $\tilde{\lambda}_t$ and any other variable of interest let n_f, n_f^{ξ} and $n_f^{\xi\tilde{\lambda}}$ denote respectively the number of bounded moments of f, the number of bounded moments of the derivative of f with respect to ξ and the number of bounded moments of the cross derivative with respect to $\xi, \tilde{\lambda}$. In the same way define $\bar{n}_f, \bar{n}_f^{\xi}$ and $\bar{n}_f^{\xi\tilde{\lambda}}$ to be respectively the number of bounded moments of the random variable $\sup_{\tilde{\lambda}} |f(\tilde{\lambda}, \xi)|$, the the number of bounded moments of the random variable $\sup_{\tilde{\lambda}} |\partial f(\tilde{\lambda}, \xi)/\partial\xi|$ and the the number of bounded moments of the random variable $\sup_{\tilde{\lambda}} |\partial^2 f(\tilde{\lambda}, \xi)/\partial\xi\partial\tilde{\lambda}_t|$.

Assumption 3.2.5.
$n_l = \min\{n_{\log \bar{g}'}, n_{\bar{p}}\} \geq 1$ and $\bar{n}_{s_t} > 0$ for all t.

This assumptions ensures the log-likelihood has bounded moments, an other standard assumption in the realm of asymptotic statistics.
The next proposition establishes the strong consistency of $\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1)$ under possible model misspecification.

Proposition 3.2.6. Blasques et al. (2014b)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary and ergodic sequence. Furthermore, let $E[|Y_t|^{n_Y}] < \infty$ for some $n_Y \geq 0$ and let assumptions 3.2.2, 3.2.3, 3.2.4 ,3.2.5 hold. Finally let $\boldsymbol{\theta}_0 \in \Theta$ be the unique maximizer of the limit log-likelihood $\tilde{\mathcal{L}}_\infty(\cdot)$ on the parameter space $\Theta \subseteq \Theta^*$. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1)$ converges almost surely to $\boldsymbol{\theta}_0$ as $T \to \infty$ for any $\bar{\lambda}_1$.

This proposition establishes the strong consistency of the maximum likelihood estimator in a possibly misspecified model setting. In particular, consistency of the maximum likelihood estimator is obtained with respect to a pseudo-true parameter that is assumed to be the unique maximum of the limit log-likelihood $\tilde{\mathcal{L}}_\infty(\boldsymbol{\theta})$. This pseudo-true parameter $\boldsymbol{\theta}_0$ minimizes the Kullback Leibler divergence between the probability measure of $\{Y_t\}_{t\in\mathbb{Z}}$ and the measure implied by the model, i. e., if the true conditional probability distribution of Y_t given y^{t-1} is $p(y_t|y^{t-1})$ then

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta}\left\{ E\left[\log \frac{p(y_t|y^{t-1})}{\tilde{p}_Y(y_t|\tilde{\lambda}_t(\boldsymbol{\theta},\bar{\lambda}_1),\xi)}\right]\right\}.$$

Convergence to the pseudo true parameter is the best we can achieve under model misspecification, see White (1982), or Monfort (1996) for a presentation that takes into account dependence in the observations.

3.2.2 Asymptotic normality

As for the case of consistency Blasques et all Blasques et al. (2014b) tackle the problem of asymptotic normality in the misspecified setting. To do this they will make use of the concept of near epoch dependence popularized in Gallant (1987) and Gallant and White (1988) that can be traced back atleast to Ibragimov (1962).

Definition 3.2.7.
A sequence of integrable random variables $\{X_t\}_{t\in\mathbb{Z}}$ on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ is called L^p near epoch dependent of size $q_0 \in \mathbb{R}$ on the stochastic process $\{Z_t\}_{n\in\mathbb{Z}}$ in $(\Omega,\mathcal{F},\mathbb{P})$ with approximation constants $\{d_i\}_{i\in\mathbb{N}}, \{v_i\}_{i\in\mathbb{N}}$, if

$$||X_i - E[X_i|\mathcal{F}_{i-m}^{i+m}]||_p \leq d_i v_m$$

where $v_m = O(m^{-q})$ for $q > q_0$ and $\mathcal{F}_{i-m}^{i+m} := \sigma(Z_{i-m},\ldots,Z_{i+m})$ is the sigma algebra generated by $Z_{i-m},\ldots,Z_{i+m}$.

Near epoch dependence is needed since in the case of misspecified models the score of the maximum likelihood needn't be a martingale difference sequence as is usually the case where no model misspecification is present. As a result, stricter conditions are required on the sequence of observations $\{Y_t\}_{t\in\mathbb{Z}}$ to obtain a central limit theorem that allows for some temporal dependence in the ML score. The more canonical strongly mixing condition does not guarantee to be strong enough of a condition since infinite distribute lag functions of strongly mixing processes are not necessarily strongly mixing, so just assuming the sequence of observations $\{Y_t\}_{t\in\mathbb{Z}}$ as being strongly mixing would not allows the asymptotic results on strongly mixing processes to be valid for all score driven models.

Definition 3.2.8.
Let $\{X_t\}_{t\in\mathbb{N}}$ be a stochastic process on a probability space $(\Omega,\mathcal{F},\mathbb{P})$. Then, the strong mixing coefficient is given by

$$\alpha(k) = \sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A\in\mathcal{F}_1^t, B\in\mathcal{F}_{t+k}^{\infty}, t\in\mathbb{N}\},$$

where $\mathcal{F}_a^l$ is the σ-field generated by $Y_a, \dots, Y_l$ and $\{X_t\}_{t\in\mathbb{N}}$ is called strongly mixing of size $-\delta$ if $\alpha(k) = \mathcal{O}(k^\beta)$ where $\beta < -\delta$.

Thus in the proof of the asymptotic normality result a convergence theorem for near epoch dependent processes due to Davidson (1992) will be used.
To obtain the result regarding asymptotic normality we define, for convenience, the following quantities utilizing the notation defined in the previous section

$$\begin{aligned}
n^* &:= \min\{n_{s_t}, \bar{n}_{s_t}^{\tilde{\lambda}_t}, \bar{n}_{s_t}^{\xi}, \bar{n}_{s_t}^{\tilde{\lambda}_t\tilde{\lambda}_t}, \bar{n}_{s_t}^{\tilde{\lambda}_t\xi}, \bar{n}_{s_t}^{\xi\xi}\} \\
n_{l'} &:= \min\left\{n_{\bar{p}}^{\xi}, \frac{n_{s_t} n_{\tilde{\lambda}_t\boldsymbol{\theta}}}{n_{s_t} + n_{\tilde{\lambda}_t\boldsymbol{\theta}}}\right\} \\
n_{l''} &:= \min\left\{n_{\bar{p}}^{\xi\xi}, \frac{n_{s_t} n_{\tilde{\lambda}_t\boldsymbol{\theta\theta}}}{n_{s_t} + n_{\tilde{\lambda}_t\boldsymbol{\theta\theta}}}, \frac{n_{s_t}^{\xi} n_{\tilde{\lambda}_t\boldsymbol{\theta}}}{n_{s_t}^{\xi} + n_{\tilde{\lambda}_t\boldsymbol{\theta}}}, \frac{n_{s_t}^{\tilde{\lambda}_t} n_{\tilde{\lambda}_t\boldsymbol{\theta}}}{2n_{s_t}^{\tilde{\lambda}_t} + n_{\tilde{\lambda}_t\boldsymbol{\theta}}}\right\}
\end{aligned}$$

We will need other assumption to obtain asymptotic normality with possible model misspecification, one on the moments and one to ensure that the score evaluated at the maximum likelihood estimator will be near epoch dependent.

Assumption 3.2.9.
There exists a set $\Theta_*^* \subseteq \mathbb{R}^4$ such that $n^* > 0, n_{l''} \geq 1$ and $n_{l'} \geq 2+\delta$ for some $\delta > 0$ and for all t.

Having $n_{l'} > 2+\delta$ facilitates the application of a central limit theorem to the score. Similarly, $n_{l''} \geq 1$ allows us to use a uniform law of large numbers for the Hessian. Finally, the condition $n^* > 0$ is designed to ensure that the e.a.s. convergence of the filter $\tilde{\lambda}_t(\boldsymbol{\theta}, \bar{\lambda}_1)$ to its stationary limit is appropriately reflected in the convergence of both the score and the Hessian.

Assumption 3.2.10.
$\partial\bar{p}_t/\partial\tilde{\lambda}_t$ and $\partial \log \bar{g}_t'/\partial\tilde{\lambda}_t$ are uniformly bounded random variables and $\partial\bar{p}_t/\partial\xi$ is almost surely Lipschitz continuous in $(Y_t, \tilde{\lambda}_t)$.

Assumption 3.2.9 imposes sufficient conditions for the score of the maximum likelihood to be Lipschitz continuous on Y_t as well as on $\tilde{\lambda}_t$. This is designed to guarantee that the score of the maximum likelihood inherits the near epoch dependence property that will be assumed on the sequence of observations $\{y_t\}_{t\in\mathbb{Z}}$.
Now we are ready to state the result on asymptotic normality in the case of model misspecification

Proposition 3.2.11. Blasques et al. (2014b)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be stationary, ergodic and near epoch dependent of size -1 on a strongly mixing process of size $-\delta/(1-\delta)$ for some $\delta > 2$. Let $E[|Y_t|^{n_Y}] < \infty$ for some $n_Y \geq 0$ and let assumptions 3.2.2, 3.2.3, 3.2.4, 3.2.5, 3.2.9 and 3.2.10 be satisfied. Furthermore let $\boldsymbol{\theta}_0 \in \operatorname{int}(\Theta)$ be the unique maximum of the log likelihood at the limit $\tilde{\mathcal{L}}_\infty(\boldsymbol{\theta})$ on Θ, where $\Theta \subseteq \Theta^* \cap \Theta^*_*$.
Then for every $\bar{\lambda}_1 \in \tilde{\Lambda}$ the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1)$ satisfies

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1) - \boldsymbol{\theta_0}) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\mathcal{J}(\boldsymbol{\theta}_0)\mathcal{I}^{-1}(\boldsymbol{\theta}_0)) \quad \text{as } T \to \infty$$

where $\mathcal{I}(\boldsymbol{\theta}_0) := E[\tilde{\mathcal{L}}''_t(\boldsymbol{\theta}_0)]$ is the Fisher information matrix, $\tilde{\mathcal{L}}_t(\boldsymbol{\theta}_0)$ denotes the log-likelihood contribution of the t-th observation evaluated at $\boldsymbol{\theta}_0$, and

$$\mathcal{J}(\boldsymbol{\theta}_0) := \lim_{T\to\infty} T^{-1} E\left(\sum_{t=1}^{T} \tilde{\mathcal{L}}'_t(\boldsymbol{\theta}_0)\right)\left(\sum_{t=1}^{T} \tilde{\mathcal{L}}'_t(\boldsymbol{\theta}_0)^\top\right)$$

When the model is correctly specified, the ML score can be shown to be a martingale difference sequence at the true parameter value. This allows to simplify the assumptions considerably and avoid the use of near epoch dependence.

3.2.3 Invertibility

Invertibility is an important property for an econometric time varying parameter model to possess if one wishes to use the model in an applied setting, here we give the definition due to Strauman and Mikosch (2006).

Definition 3.2.12. Strauman and Mikosch (2006)
Given a $\boldsymbol{\theta} \in \Theta$ we say that a correctly specified observation driven model 3.1 is invertible if

$$||\tilde{\lambda}_t - \lambda_t|| \xrightarrow{\mathbb{P}} 0$$

for any $\bar{\lambda}_1 \in \tilde{\Lambda}$.

Invertibility tells us that, given a vector of parameters, no matter the starting value $\bar{\lambda}_1$ from which we choose to initialize the model we will converge to the the true time varying parameter. Of course of great relevance is the case when the vector of fixed parameters are the true ones $\boldsymbol{\theta}_0$, then, given an arbitrary large sample from the process $\{y_t\}_{t=1}^T$, the arbitrary choice of the model initialization will have negligible impact on

the goodness of the model. There exist in fact cases where knowing the true parameter vector $\boldsymbol{\theta}_0$, even under correct model specification, does not guarantee the true time varying parameter will be tracked by the model if not initialized properly, see Sorokin (2011), where some GARCH-type models are shown to admit a stationary solution but lack invertibility. This problem has also been discussed in Wintenberger (2013).
Another important reason for requiring model invertibility is that the the log-likelihood function $\tilde{\mathcal{L}}_T(\boldsymbol{\theta}, \bar{\lambda}_1)$ depends on the initial condition $\bar{\lambda}_1$ thus in turn the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_T(\bar{\lambda}_1)$ may also depend on $\bar{\lambda}_1$. This poses a problem for establishing consistency and asymptotic normality since there is no guarantee that asymptotically this dependency will vanish, thus an unwanted degree of arbitrariness given by the choice of $\bar{\lambda}_1$ may be present even with arbitrarily large sample sizes. For this issue it is not enough that the model be invertible at the true parameter value $\boldsymbol{\theta}_0$ since the log-likelihood function is maximized over the entire set Θ, thus another invertibility definition was introduced in the modern literature by Wintenberger (2013)

Definition 3.2.13. Wintenberger (2013)
A correctly specified observation driven model 3.1 is continuously invertible on a compact set Θ if

$$\sup_{\boldsymbol{\theta}\in\Theta} ||\tilde{\lambda}_t - \lambda_t|| \xrightarrow{\mathbb{P}} 0$$

for any $\bar{\lambda}_1 \in \tilde{\Lambda}$.

Continuous invertibility allows the asymptotics of the maximum likelihood estimator to be independent from the initial condition $\bar{\lambda}_1$ allowing for the possibility of proving consistency and asymptotic normality. Indeed in Blasques et al. (2014b) model continuous invertibility is assumed through conditions (i) and (ii) of theorem 3.2.1, a theorem that is one of the main building blocks to demonstrate their results on consistency and asymptotic normality.
In the case the model is not correctly specified, that is when the assumed probability density $\tilde{p}$ is not equal to the true density p of the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$ a more realistic notion of invertibility is needed, since if we misspecified the model it will be unlikely we will manage to converge to the true time varying parameter just as when the data has no dependence structure, see White (1982).
In the case of time varying parameter models there are two types of misspecification: one coming from erroneously specifying the distribution $\tilde{p}_t$ as not equal to the one of Y_t, the other coming from choosing the dynamics of $\{\tilde{\lambda}_t\}_{t\in\mathbb{Z}}$ in a way that does not cover the evolution of the true time varying parameter λ_t through time. It is well known that

in the case of model misspecification the maximum likelihood estimator is a natural estimator for the vector of parameters that minimize the Kullback Leibler (KL) divergence between the true probability density and the model density, Akaike (1998). Thus calling λ_t^* the time varying parameter that minimizes the KL divergence between the true density $p_t(y_t|\lambda_t)$ at time t (the true density could even be non parametric) and the model density $\tilde{p}_t(y_t|\tilde{\lambda}_t)$ (notice that we do not take into account the specified dynamics of $\tilde{\lambda}_t$ but only the shape of the distribution, so λ_t^* does not depend on $\boldsymbol{\theta}$) we give the following definition

Definition 3.2.14.
Given a $\boldsymbol{\theta} \in \Theta$ we say that an observation driven model 3.1 is invertible if

$$||\tilde{\lambda}_t - \lambda_t^*|| \xrightarrow{\mathbb{P}} 0$$

for any $\bar{\lambda}_1 \in \tilde{\Lambda}$.

Section 4.2.3 will give indications that score driven models posses properties that may imply this definition of invertibility thanks to the score present in the dynamic equation.

Chapter 4

Optimality of Score Driven Models

The use of the score in the updating equation of observation driven models has lead to the specification of models that perform well in an applied setting. Due to the relatively recent specification of the class of score driven models, the entire set of mathematical properties, that derive from the use of the derivative of the log-likelihood in the dynamic equation, has not yet been completely uncovered. The following section presents the current literature on the argument and subsequently derives novel mathematical properties for a (prototypical) class of score driven models. Indeed, understanding in which cases score driven models are well suited to model a time series with a dependence structure is valuable information for practitioners that have to choose the type of model on the basis of the available data.
All proofs in this chapter can be found in appendix B

4.1 A Review of the Current Literature

Recently, Blasques et al. (2015), proved an information theoretical criterion for a specific score driven updating equation of the first order, in the one dimensional case.
In the paper, a general family of observation driven models is considered under the framework of model misspecification. Formally, it is assumed that the data follows a real valued discrete time stochastic process $\{Y_t\}_{t\in\mathbb{N}}$, the data generating process (DGP), that takes values in $\mathcal{Y} \subseteq \mathbb{R}$ and has probability density $p(y_t|\lambda_t)$ at time t, where $\lambda_t \in \mathbb{R}$ for all t. Having observed a given sequence of T observations $\{y_t\}_{t=1}^{T}$ from the DGP the authors consider a generic first order observation driven model

$$\begin{aligned} &Y_t \sim \tilde{p}(y_t|\tilde{\lambda}_t, \boldsymbol{\theta}),\\ &\tilde{\lambda}_{t+1} = \phi(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) \end{aligned} \tag{4.1}$$

where $\tilde{p}(\cdot|\tilde{\lambda}_t, \boldsymbol{\theta})$ is a parametric density, $\tilde{\lambda}_t \in \tilde{\Lambda} \subset \mathbb{R}$ is a filtered value of the true λ_t that may depend on the observations up to time t, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ is a vector of static unknown parameters, and ϕ is an updating function linking the one step ahead time varying parameter $\tilde{\lambda}_{t+1}$ to the current observation y_t and the current filtered time-varying parameter $\tilde{\lambda}_t$. To simplify the notation, when there is no possibility of confusion, p_t will be used to indicate $p(y|\lambda_t)$ and $\tilde{p}_t$ or $\tilde{p}(y|\tilde{\lambda}_t)$ to mean $\tilde{p}(y|\tilde{\lambda}_t, \theta)$).

The assumption of possible model misspecification, i.e., $p(y_t|\lambda_t)$ is not necessarily equal to $\tilde{p}(y_t|\tilde{\lambda}_t, \boldsymbol{\theta})$, keeps the analysis as realistic and general as possible. The choice of the updating function ϕ is usually made so that the model approximates λ_t better and better as more realizations of the process are observed.

It is then argued that an ideal property for any observation driven model to have is that the sequence of values $\{\tilde{\lambda}_t\}$ they generate obey an optimal information theoretic update, i.e., the update from $\tilde{\lambda}_t$ to $\tilde{\lambda}_{t+1}$ decreases the Kullback-Leibler (KL) divergence between the true conditional density $p(\cdot|\lambda_t)$ and the model postulated density $\tilde{p}(\cdot|\tilde{\lambda}_t; \theta)$. The KL distance under consideration is given by

$$D_{KL}(p_t, \tilde{p}_t) = \int_A p(y|\lambda_t) \ln \frac{p(y|\lambda_t)}{\tilde{p}(y|\tilde{\lambda}_t; \theta)}\, dy$$

where $A \subset \mathbb{R}$ is the subset of the real line over which the divergence is evaluated.

In particular, given a starting point for the filtered parameter $\tilde{\lambda}_t$ and the true density $p(\cdot, \tilde{\lambda}_t)$ the authors analyze the conditions under which a new observation y_t, drawn from the true conditional density of the DGP $p(\cdot, \lambda_t)$, produces an update from $\tilde{\lambda}_t$ to $\tilde{\lambda}_{t+1}$ such that the new conditional density $\tilde{p}(\cdot|\tilde{\lambda}_{t+1})$ provides a better approximation to $p(\cdot|\lambda_t)$ than $\tilde{p}(\cdot|\tilde{\lambda}_t)$.

There remains a caveat: in an applied setting it is mathematically implausible that $\tilde{p}(\cdot|\tilde{\lambda}_{t+1})$ approximates $p(\cdot|\lambda_{t+1})$ better than $\tilde{p}(\cdot|\tilde{\lambda}_t)$ approximates $p(\cdot|\lambda_t)$. The problem is that $\tilde{\lambda}_{t+1}$ is updated using information from the density of the DGP at the previous time $p(\cdot|\lambda_t)$ and therefore, without imposing any restriction on the true sequence of conditional densities, it is impossible to say if the updating scheme approximates the one step ahead density of the DGP $p(\cdot|\lambda_{t+1})$. For this reason the optimality criterion is computed with respect to $\tilde{p}(\cdot|\tilde{\lambda}_{t+1})$ and the true density at the previous time $p(\cdot|\lambda_t)$.

Definition 4.1.1. Blasques et al. (2015)
The Realized Kullback Leibler (RKL) variation of a parameter update from $\tilde{\lambda}_t \in \tilde{\Lambda} \subseteq R$ to $\tilde{\lambda}_{t+1} \in \tilde{\Lambda} \subseteq R$ is given by

$$\begin{aligned}\Delta_{t|t} &= D_{KL}(p_t, \tilde{p}_{t+1}) - D_{KL}(p_t, \tilde{p}_t)\\ &= \int_A p(y|\lambda_t)(\ln \tilde{p}(y|\tilde{\lambda}_t; \theta) - \ln \tilde{p}(y|\tilde{\lambda}_{t+1}; \theta))dy\end{aligned}$$

Where $A \subseteq \mathbb{R}$ is the subset of the real line over which the divergence is evaluated. For a given p_t, a parameter update will be said to be RKL optimal if and only if $\Delta_{t|t} < 0$

Thus, if the RKL variation is negative, then $\tilde{p}(\cdot|\tilde{\lambda}_{t+1})$ approximates $p(\cdot|\lambda_t)$ better than $\tilde{p}(\cdot|\tilde{\lambda}_t)$.

The authors in their main proposition show that only particular first order observation driven models will be RKL optimal, specifically ones that utilize the score in the updating equation for $\tilde{\lambda}_t$. They are defined as follows.

Definition 4.1.2. Newton-score update Blasques et al. (2015)
A Newton-score update model is defined as

$$\begin{aligned} &Y_t \sim \tilde{p}(y_t|\tilde{\lambda}_t, \boldsymbol{\theta}),\\ &\tilde{\lambda}_{t+1} = \phi(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) := \tilde{\lambda}_t + \alpha S_t \tilde{s}_t,\\ &\tilde{s}_t := \tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) := \partial \log \tilde{p}(y_t|\lambda, \boldsymbol{\theta})/\partial\lambda|_{\lambda=\tilde{\lambda}_t} \end{aligned} \tag{4.2}$$

where $S_t := S(\tilde{\lambda}_t, \boldsymbol{\theta})$ is a positive scaling function that possibly depends on the filtered time varying parameter $\tilde{\lambda}_t$ and the static parameter $\boldsymbol{\theta}$.

As one can notice, the Newton-score update is a score driven model with a simple dynamic equation for the time varying parameter that depends only on a single static parameter: α.

The conditions that are assumed for the main proposition of Blasques et al. (2015) to hold are regularity conditions on the support of the true conditional $p(y_t|\lambda_t)$, the score of the postulated model $\tilde{s}_t$ should be non zero, and a crucial sign condition on both α and S_t.

Assumption 4.1.3.
$p(y_t|\lambda) > 0 \;\forall\; (y_t, \lambda) \in \mathbb{R} \times \Lambda$ and $\tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) \neq 0$ for every $(\tilde{\lambda}_t, \boldsymbol{\theta}) \in \tilde{\Lambda} \times \Theta$ and almost every $y_t \in \mathbb{R}$ for all t.

Assumption 4.1.4. *$\alpha > 0$ and $S(\tilde{\lambda}_t, \boldsymbol{\theta}) > 0 \;\forall\; (\tilde{\lambda}_t, \boldsymbol{\theta}) \in \tilde{\Lambda} \times \Theta$ and all t.*

The first statement in 4.1.3 excludes those values of the time-varying parameter that result in a distribution for Y_t that is degenerate while the second statement in 4.1.3

excludes the possibility of the time-varying parameter being non-identified at certain update steps. Condition 4.1.4 instead imposes two sign conditions on α and $S(\tilde{\lambda}_t, \boldsymbol{\theta})$ to ensure the information contained in the score is not "distorted", technically, as will be made clear in section 4.2, since the score indicates the direction of steepest ascent towards the log-likelihood maximum one needs to ensure that the sign of the score is retained in the model specification.

Proposition 4.1.5. Blasques et al. (2015)
Under Condition 4.1.3 and 4.1.4 every Newton-score update is locally RKL optimal for any true density p_t. Where the word locally signifies that $\exists\ \delta_y, \delta_\lambda > 0$ such that $\Delta_{t|t} < 0$ for all sets of the form

$$\Lambda_{\delta_\lambda}(\tilde{\lambda}_t) := \{\tilde{\lambda} \in \tilde{\Lambda} : |\tilde{\lambda} - \tilde{\lambda}_t| < \delta_\lambda\}$$
$$Y_{\delta_y}(y_t) := \{y \in \mathcal{Y} : |y - y_t| < \delta_y\}$$

The proof is given in appendix B.

An issue with the result, as reformulated in Blasques et al. (2018), lies in the local nature of the derivations. Indeed, locally, the KL divergence can be negative. The KL divergence is necessarily positive only if evaluated over the entire support of the probability densities. In turn this leads to a breakdown in the significance of $\Delta_{t|t}$ since imposing that

$$D_{KL}(p_t, \tilde{p}_{t+1}) - D_{KL}(p_t, \tilde{p}_t) < 0$$

locally does not guarantee that we are approximating p_t better with $\tilde{p}_{t+1}$ than $\tilde{p}_t$.

Hence, a sufficient condition for the local KL divergence to be positive is that the set Y_{δ_y} defined as a neighborhood of y_t, must be such that the true conditional density p_t dominates the filtered density $\tilde{p}_t$, i.e., $p_t(y) > \tilde{p}_t(y)$ for all $y \in Y_{\delta_y}$. The revision given in Blasques et al. (2018) proceeds to state that "in general, given the local nature of the Y_{δ_y}, the majority of the realizations y_t will deliver a set Y_{δ_y} that satisfies the additional condition."

Recently there have been other attempts at proving why score driven models are optimal among observation driven models: an extensive Monte Carlo study was performed in Blasques et al. (2017) where the RKL divergence of the GARCH model, the fat-tailed t-GARCH model of Bollerslev (1987) , the t-GAS model of Creal et al. (2011) and the log-GAS model introduced by Harvey (2013) was computed. Score driven models were found to perform well, in this finite sample setting, especially when the data used was fat-tailed. In Blasques et al. (2019) score driven models were compared to other well-known nonlinear dynamic models such as the threshold model Tong and Lim (2009) and smooth transition auto-regressive model Chan and Tong (1986) and found to perform well.

4.2 A Novel Optimality Criterion

In this section, motivated by the framework developed in Blasques et al. (2015), we prove that score driven models possess an intuitive, global and high-dimensional property.
The setting will be the one of full generality discussed up until now, with the difference that we focus on the sequence of pseudo-true parameters $\{\lambda_t^*\}_{t\in\mathbb{Z}}$, i.e., the sequence of values that maximize the sequence of expected pseudo-true likelihoods

$$\lambda_t^* = \arg\max_{\tilde{\lambda}} E_{Y_t}[\log \tilde{p}(Y_t|\tilde{\lambda}, \theta)]$$

since these are the values that minimize the Kullback Leiber divergence between the true probability density of the DGP and the model density at each time t, as discussed in the seminal papers by Akaike (1998) and White (1982). We also remark, to avoid confusion, that we will now consider score driven models as stochastic processes rather than as realizations of stochastic processes, i.e., we utilize definition 3.2 instead of 4.1.
The tuple formed by the expectation of the pseudo-true likelihood viewed as a function of the time varying parameter of interest $f_t(\tilde{\lambda})$ and the sequence of pseudo-true parameters, i.e. ,

$$f_t(\tilde{\lambda}) : \tilde{\lambda} \to E_{Y_t}[\log \tilde{p}(Y_t|\tilde{\lambda}, \theta)], \qquad \{\lambda_t^*\}_{t\in\mathbb{Z}}$$

will be the main building blocks of our analysis.
Prompted by the framework discussed in section 4.1 we focus on the improvement that the score driven updating step produces in terms of distance from the pseudo-true time varying parameter λ_t^*. The novelty here is that, rather than looking at a notion of distance between probability distributions, as was previously given by the KL divergence, we will focus on the Euclidean distance between the two pseudo-true time varying parameters $\lambda_t^*, \lambda_{t+1}^*$ and the model parameter $\tilde{\lambda}_t$ at each time t. As was the case in Blasques et al. (2015), it is not mathematically feasible to always decrease the distance to the pseudo-true parameter λ_{t+1}^* with an observation y_t that comes from the true density at the previous time $p(y_t|\lambda_t)$ unless we introduce assumptions on the evolution through time of the pseudo-true time varying parameter. Imposing conditions on the evolution of the pseudo-true time varying parameter will be discussed in section 4.2.2, where some results are given in this direction.
Specifically, given an initial value for the model parameter $\tilde{\lambda}_t$, we prove, under some conditions, that the new observation y_t, drawn from $p(y_t|\lambda_t)$, produces an update from $\tilde{\lambda}_t$ to $\tilde{\lambda}_{t-1}$ that always decreases the Euclidean distance from the pseudo-true parameter λ_t^* in

conditional expectation. The decrease, the expectation, of the Euclidean distance will be proved by solving, through techniques belonging to the stochastic gradient descent theory, the aforementioned time varying optimization problem specified by the sequence of model expected log-likelihoods. We then give a definition for such a property that formalizes the fact already noted by Creal et al. (2013) that the score points in the "steepest ascent direction for improving the model's local fit in terms of the likelihood or density at time t". Indeed the properties is a natural consequence of the monotonicity in expectation of the stochastic gradient descent algorithm 2.3 at each time step t. The conditional expected variation optimality property thus mirrors the notion of KL optimality on the parameter space instead of on the space of distributions. We also explicitly remark that the expectation notation $E_{Y_t|Y^{t-1}}[\cdot]$ indicates the conditional expectation of Y_t with respect to all previous random variables $Y_{t-1}, Y_{t-2}, \dots$.

Definition 4.2.1. Conditional Expected Variation Optimality
The conditional expected variation (CEV) of a parameter update from $\tilde{\lambda}_t \in \tilde{\Lambda} \subseteq \mathbb{R}$ to $\tilde{\lambda}_{t+1} \in \tilde{\Lambda} \subseteq R$ at time t is given by

$$\Delta_{E_{t|t-1}} = \left\| \lambda_t^* - E_{Y_t|Y^{t-1}}\left[\tilde{\lambda}_{t+1}\right] \right\| - \left\| \lambda_t^* - \tilde{\lambda}_t \right\|$$

a parameter update will be said to be CEV optimal if and only if $\Delta_{E_{t|t-1}} < 0$

Besides simplifying the analysis and making it more intuitive the reason for this change of distance from the space of distributions to the parameter space comes mainly from the fact that, even in the case of correct model specification, if we get closer to the true time varying parameter it does not imply that we get closer to the true distribution in KL divergence. That is, $\left\| \lambda_t - \tilde{\lambda}_{t+1} \right\| \leq \left\| \lambda_t - \tilde{\lambda}_t \right\|$ does not imply

$$D_{KL}(p_t(y_t|\tilde{\lambda}_{t+1}), p_t(y_t|\lambda_t)) \leq D_{KL}(p_t(y_t|\tilde{\lambda}_t), p_t(y_t|\lambda_t))$$

And the dynamic equation 3.2 aims to predict the time varying parameter not the time varying distribution. Because of this fact, and the local nature of the RKL variation, it is hard to establish a clear relationship between the two optimality concepts.

If CEV optimality holds for all t then it is also closely related to the desirable invertibility property 3.2.14 since then the model time varying parameter gets closer (in expectation) at each step to the pseudo-true time varying parameter no matter what initial value is given to start the recursion. We discuss this more thoroughly in section 4.2.3.

As before to simplify the notation we will use unambiguously the symbols $\tilde{p}(y_t|\tilde{\lambda}_t)$ or $\tilde{p}_t$ to represent $\tilde{p}(y_t|\tilde{\lambda}, \theta)$ and p_t to represent $p_t(y_t|\lambda_t)$.

We also take for granted the following standard assumption that holds true for most model probability densities $\tilde{p}(y_t|\tilde{\lambda}_t)$

Assumption 4.2.2.
The gradient with respect to the parameter of interest $\tilde{\lambda}$ can be exchanged with respect to the conditional expectation of the stochastic process at any time t, i.e.

$$\nabla E_{Y_t|Y^{t-1}}[-\log \tilde{p}(Y_t|\tilde{\lambda},\theta)] = E_{Y_t|Y^{t-1}}\left[-\nabla \log \tilde{p}(Y_t|\tilde{\lambda},\theta)\right] \quad \forall t.$$

With this assumption it is immediately possible to state the conditional expected variation optimality for a Newton-score update model 4.2, that mirrors the result in Blasques et al. (2015). The proof will be given in the respective appendix as all the others to come and checking the assumptions on the sequence $\{f_t(\tilde{\lambda})\}_{t\in\mathbb{Z}}$ in an applied setting will be discussed in section 4.2.1.

Proposition 4.2.3. CEV Optimality
Let assumptions 2.0.1,2.0.2 and 4.2.2 hold for every function of the sequence $\{f_t(\tilde{\lambda})\}_{t\in\mathbb{Z}}$. Then every Newton-score update 4.2 with $0 < S_t\alpha < 2/L$ $\forall\tilde{\lambda}_t$ is CEV optimal for any true density p_t, i. e.,

$$\Delta_{E_{t|t-1}} < 0 \qquad \forall t$$

This results is chiefly theoretical since in most practical applications the sequence $\{f_t(\tilde{\lambda})\}_{t\in\mathbb{Z}}$ is not known (because the expectation is taken with respect to the true distribution that is usually unknown) so the conditions appear hard to check. Nonetheless, for the purpose of this thesis, the proposition holds great pedagogical value since it reveals the connection to optimization theory and will be introductory to the ideas behind the extensions in the next sections.
As in the proof of proposition 4.1.5 the crucial step is to recover the square of the score that gives a crucial sign information used throughout the derivations. While, in contrast to proposition 4.1.5, the proof outline comes directly from the usual machinery of optimization theory, as can be seen the standard inequalities are used, there is no need for a local argument and the results hold globally. The reason one needs the extra condition $S_t\alpha < 2/L$ in addition to $S_t\alpha > 0$ come exactly from the global nature of the result (generalizations of this result will impose greater conditions on S_t highlighting its role as can bee seen in section 4.2.1). Intuitively the reason why the Lipschitz assumption is not needed in proposition 4.1.5 is that one can always find a small enough neighborhood at any point such that a continuous function in that neighborhood will have Lipschitz continuous gradient. Moreover, as can be seen in the proof of the result, one could actually show convergence in expectation of the model time varying parameter $\tilde{\lambda}_t$ to a neighborhood of the pseudo-true time varying parameter λ_t^* at a fixed time t if it were

possible to have infinite extractions of the random variable Y_t. The implication of this point of contact between the optimization theory and the econometric theory regarding score driven models will be the centerpiece of this thesis.

Thus it will be instrumental, to understand the intuition behind proposition 4.2.3 and the other upcoming propositions, to recast the problem of predicting the time varying parameter as a time varying optimization problem. On one hand, in the econometric literature, when modeling a time series with a time varying parameter we proceed as follows: Having observed a given sequence of T observations $\{y_t\}_{t=1}^T$ we choose a model, with the objective of tracking the time varying parameter as closely as possible, that we believe could fit the data well. Then we estimate the parameters and can perform various diagnostics for model correctness. The goal is for the modeled time varying parameter $\tilde{\lambda}_t$ to be as close as possible to λ_t^*.

On the other hand, under some assumptions on the probability density, λ_t^* is exactly the maximum of the function

$$f_t(\lambda) : \lambda \rightarrow -E[\log p(Y_t|\lambda)]$$

this results in a sequence of maxima $\{\lambda_t^*\}_{t\in\mathbb{Z}}$ each one associated with the corresponding objective function $f_t(\lambda)$ and the problem of tracking this sequence of maxima is addressed by the time varying optimization theory. In this sense time varying optimization of the sequence of expected model log likelihoods and observation driven modelling of a time varying parameter share the same objective. This thesis argues that score driven models are a point of contact between the econometric theory and the time varying optimization theory. Having already anticipated how the two theories, in the time varying parameter case, share the same goals, we will prove in following sections some extensions of the result of CEV optimality that will require weaker assumptions. To do so we will utilize techniques (as in proposition 4.2.3) that are characteristic of the optimization framework.

4.2.1 Conditional Expected Variation Optimality in Practice

In this section we analyze CEV optimality of score driven models and generalize it further rendering it obtainable in applied settings.

First we wish to find an analog of proposition 4.2.3 that has assumptions that can be checked in practice, where we don't know the true distribution.

Again our main assumptions will be based on the sequence of objective functions (to emphasize the connection with the optimization literature we will use the optimization framework nomenclature and call the sequence of functions to be minimized the sequence of objective functions) that is, as already anticipated, the sequence of expected log likelihoods

$$\{f_t(\tilde{\lambda})\}_{t\in\mathbb{N}} := \{-E_{Y_t|Y^{t-1}}[\log \tilde{p}(Y_t|\tilde{\lambda}, \boldsymbol{\theta})]\}_{t\in\mathbb{Z}}$$

where one can recognize $\log \tilde{p}(Y_t|\tilde{\lambda}, \boldsymbol{\theta})$ as the pseudo log-likelihood function of Y_t. Notice that the conditional Expectation is used because, for score driven models, $\tilde{\lambda}$ is a function of past random variables, i. e., $\tilde{\lambda} := \tilde{\lambda}(Y^{t-1})$ but often we will want to treat it like a constant. While the reason we utilize a minus sign in the sequence of objective functions is to abide with the convention of the optimization literature to talk about minimization rather than maximization, of course the problems are equivalent.

In this context the minimum of the objective function $f_t(\tilde{\lambda})$ is the pseudo-true time varying parameter λ_t^* that minimizes the KL divergence between the true density p_t and the model one $\tilde{p}_t$. This, as already stated, was highlighted by Akaike (1998), and cited by White (1982). A simple way to see this is given by the equality

$$E_{Y_t|Y^{t-1}}[\log \tilde{p}(Y_t|\tilde{\lambda}, \boldsymbol{\theta})] = E_{Y_t|Y^{t-1}}[\log p(Y_t|\lambda_t)] - E_{Y_t|Y^{t-1}}\left[\log \frac{p(Y_t|\lambda_t)}{\tilde{p}(Y_t|\tilde{\lambda}, \boldsymbol{\theta})}\right]$$

the term $E_{Y_t|Y^{t-1}}[\log p(Y_t|\lambda_t)]$ does not depend on $\tilde{\lambda}$ so to minimize $f_t(\tilde{\lambda})$ one needs to minimize the KL divergence between the true density and the model one.

So the corresponding sequence of points where the sequence of objective functions assumes a minimum is nothing else than the sequence of pseudo-true parameter values $\{\lambda_t^*\}_{t\in\mathbb{N}}$ that we wish to track as closely as possible. Specifically we would like each function of the sequence $\{f_t(\tilde{\lambda})\}_{t\in\mathbb{Z}}$ to be under assumptions 2.0.1 and 2.0.2 stated in chapter 2. Alas in applications we don't usually know this sequence because the expectation is with respect to the true density of Y_t which we generally have no access to. So our new assumptions must involve only the model density $\tilde{p}(y_t|\tilde{\lambda}, \boldsymbol{\theta})$ and the choice of S_t in the updating equation 4.2.

Assumption 4.2.4.

Let the model density be twice differentiable $\tilde{p}(y_t|\tilde{\lambda}, \boldsymbol{\theta})$ with respect to $\tilde{\lambda}$ and have unique global maximum λ_t^*. Also let

$$\nabla S_t\tilde{s}_t := -\nabla \left(S_t(\tilde{\lambda}, \boldsymbol{\theta})\nabla \log \tilde{p}(y_t|\tilde{\lambda}, \boldsymbol{\theta})\right) \leq LI$$

for all $\tilde{\lambda} \in \tilde{\Lambda}, y_t \in \mathbb{R}$ and all t, where $L \in \mathbb{R}^+$.

Assumption 4.2.5.

Let the model density be twice differentiable $\tilde{p}(y_t|\tilde{\lambda}, \boldsymbol{\theta})$ with respect to $\tilde{\lambda}$ and have unique global maximum λ_t^*. Also let

$$0 < \ell I \leq \nabla S_t\tilde{s}_t$$

for all $\tilde{\lambda} \in \tilde{\Lambda}, y_t \in \mathbb{R}$ and all t, where $\ell \in \mathbb{R}^+$.

With these conditions on the model density $\tilde{p}(y_t|\tilde{\lambda}, \boldsymbol{\theta})$ and on $S_t(\tilde{\lambda}_t, \boldsymbol{\theta})$ we will prove proposition 4.2.7 that utilizes lemma 2.2.2 as inspiration in the stochastic time varying gradient descent setting. As in lemma 2.2.2 one can make assumptions 4.2.4 and 4.2.5 hold by choosing accurately S_t so as to modify the score of the model density at each time t making it a Lipschitz continuous gradient of a strongly convex function. But the sequence of objective functions that we wish to optimize involves also the expectation with respect to the unknown density p_t, the next lemma tells us that this is not a problem.

Lemma 4.2.6.
Let assumptions 4.2.2, 4.2.4, 4.2.5 hold. Then

$$0 < \ell I \leq -\nabla E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] \leq IL$$

for all t and λ_t^* will uniquely satisfy $E_{Y_t}[S_t(\lambda_t^*, \boldsymbol{\theta}) \nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] = 0$.

Although there is no mention of the true density p_t in neither of the two conditions 4.2.4, 4.2.5 we have still managed to bound the gradient of the conditional expectation of $S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})$.

Lemma 4.2.6 tells us that $E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})]$ behaves like a Lipschitz continuous gradient of a strictly convex function that has minimum in λ_t^* mirroring lemma 2.2.2 that required the same inequalities as assumptions in the deterministic case. Notice also that one does not need to know λ_t^* but merely that it exists, this is often an implicit assumption used when finding estimators through maximum likelihood; that the log-likelihood does indeed posses a maximum. Moreover the existence and uniqueness of λ_t^* is a direct implication of the strong convexity of the sequence of model densities for all $y_t \in \mathbb{R}$ since this implies the strong convexity on the sequence of objective functions, i.e., assumption 4.2.5 and assumption 4.2.2 imply that the value that would minimize the expected value of the negative log-likelihood at time t would be the uniquely determined pseudo-true parameter value λ_t^*.

Now we can give another proposition that is mostly just a restatement of proposition 4.2.3.

Proposition 4.2.7.
Let assumptions 4.2.4, 4.2.5 and 4.2.2 hold then if $0 < \alpha < 2/L$ the Newton-score update 4.2 is CEV optimal for any true density p_t, i. e.,

$$\Delta_{E_{t|t-1}} < 0 \qquad \forall t$$

Thus for the optimality property stated in definition 4.2.1 to hold one has to choose the model density $\tilde{p}_t$ and S_t accurately. Usually this means that the sequence of objective functions $f_t(\tilde{\lambda})$ are convex and have a unique minima but they lack a Lipschitz continuous gradient. In such a case, choosing S_t so that assumptions 4.2.4 and 4.2.5 hold, one still manages to obtain CEV optimality as will be seen in the upcoming examples.
This is the first time, to the authors knowledge, that a prescription for S_t is made based on a theoretical result and not a rule of thumb. Since assumption 4.2.4 and 4.2.5 can be checked comfortably in an applied setting (the model distribution is a choice of the statistician) and there is essentially no dependence on the true density that is usually unknown (only assumption 4.2.2 is required that is a standard assumption) this proposition, as opposed to proposition 4.2.3, holds a more practical value.
The curious reader could wonder how restrictive are assumptions 4.2.4 and 4.2.5 and if they allows for variety of selections when choosing the model density, besides they require the inequality to hold uniformly in $y_t \in R$ and $\tilde{\lambda}_t \in \tilde{\Lambda}$, a feat that does not seem easy. In practice the choice of the time varying parameter of interest, of the model density, of S_t and even of the link function often used in observation driven models allow these conditions to hold in a plethora of cases. Here we give some examples for which the conditions hold

Example 4.2.8.
Let the model density $\tilde{p}_t$ be chosen as Gaussian for all t and let the parameter of interest be the mean $\tilde{\mu}$ then

$$\nabla \log \tilde{p}_t(y_t|\tilde{\mu}, \sigma^2) = \frac{\partial}{\partial \tilde{\mu}} \left(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_t - \tilde{\mu})^2}{2\sigma^2} \right) = (y_t - \tilde{\mu})/\sigma^2$$

So simply choosing $S_t(\tilde{\mu}, \sigma^2) = 1$ we recover assumptions 4.2.4 and 4.2.5 given that

$$-\nabla \left(S_t(\tilde{\mu}, \sigma^2) \nabla \log \tilde{p}_t(y_t|\tilde{\mu}, \sigma^2) \right) = 1/\sigma^2$$

As usual the likelihood for the mean of a Gaussian distribution behaves very nicely and allow us to not bring into play $S_t(\tilde{\mu}, \sigma^2)$.

Example 4.2.9.
Let the model density $\tilde{p}_t$ be chosen as Gaussian for all t and let the parameter of interest be the variance $\tilde{\sigma}^2$ then

$$\nabla \log \tilde{p}_t(y_t|\tilde{\sigma}^2, \mu) = \frac{\partial}{\partial \tilde{\sigma}^2} \left(\log \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} - \frac{(y_t - \mu)^2}{2\tilde{\sigma}^2} \right) = -\frac{1}{2\tilde{\sigma}^2} + \frac{(y_t - \mu)^2}{2\tilde{\sigma}^4}$$

Choosing $S_t(\tilde{\sigma}^2, \mu) = \tilde{\sigma}^4$ we recover assumptions 4.2.4 and 4.2.5 given that

$$-\nabla \left(S_t(\tilde{\sigma}^2, \mu) \nabla \log \tilde{p}_t(y_t | \tilde{\sigma}^2, \mu) \right) = \frac{1}{2}$$

The natural choice of $S_t(\tilde{\sigma}^2, \mu)$ in this setting is analogous to the choice of the inverse of the information matrix a proposal that was already given in the literature Creal et al. (2013) as a rule of thumb, this will be discussed more deeply in section 4.2.2 .

Example 4.2.10.
Let the model density $\tilde{p}_t$ be chosen as exponential for all t and let the parameter of interest be the scale parameter $\tilde{\beta}$, i.e.,

$$\tilde{p}_t(y_t | \tilde{\beta}, \mu) = \frac{1}{\tilde{\beta}} \exp(-y_t / \tilde{\beta}), \quad \tilde{\beta} > 0$$

then

$$\nabla \log \tilde{p}_t(y_t | \tilde{\beta}, \mu) = -\frac{1}{\tilde{\beta}} + \frac{y_t}{\tilde{\beta}^2}$$

Choosing $S_t(\tilde{\beta}) = \tilde{\beta}^2$ we recover assumptions 4.2.4 and 4.2.5 given that

$$-\nabla \left(S_t(\tilde{\beta}) \nabla \log \tilde{p}_t(y_t | \tilde{\beta}, \mu) \right) = 1$$

One can check that if we had chosen as our time varying parameter $\tilde{\lambda}$ (the one usually used to parametrize the exponential)

$$\tilde{p}_t(y_t | \tilde{\lambda}, \mu) = \tilde{\lambda} \exp(-y_t \tilde{\lambda}), \quad \tilde{\lambda} > 0$$

it would have been impossible to recover assumptions 4.2.4 and 4.2.5 given that

$$\nabla \log \tilde{p}_t(y_t | \tilde{\lambda}, \mu) = \frac{1}{\tilde{\lambda}} - y_t$$

and this gradient is not amendable to corrections through the choice of S_t.

As we can see to satisfy assumptions 4.2.4 and 4.2.5 up until now we have essentially used two tools: one is the appropriate choice of the time varying parameter the other is the choice of the function $S_t(\tilde{\lambda}, \boldsymbol{\theta})$.
However it might not always be possible to recover a Lipschitz continuous gradient of a strongly convex function but strict convexity will often be achievable as the next examples will showcase.

Example 4.2.11.
Let the model density $\tilde{p}_t$ be chosen as generalized student-t and let the parameter of interest be the scale $\tilde{\sigma} > 0$ i.e.,

$$\tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\tilde{\sigma}} \left(1 + \frac{1}{\nu}\left(\frac{y_t - \mu}{\tilde{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}}$$

Then

$$\nabla \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu) = -\frac{1}{\tilde{\sigma}} + (\nu + 1)\left(\frac{(y_t - \mu)^2}{\nu\tilde{\sigma}^3} \Big/ 1 + \frac{1}{\nu}\left(\frac{y_t - \mu}{\tilde{\sigma}}\right)^2\right)$$

choosing $S_t(\tilde{\sigma}) = \tilde{\sigma}$ yields

$$-\nabla\left(S_t(\tilde{\sigma})\nabla \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu)\right) = \frac{(\nu + 1)(y_t - \mu)^2 2\nu\tilde{\sigma}}{(\nu\tilde{\sigma}^2 + (y_t - \mu)^2)^2}$$

that is strictly positive for all $\tilde{\sigma}, y_t$ and is also jointly bounded above for all $\tilde{\sigma}, y_t$, this can be seen by taking the limits as $\tilde{\sigma}, y_t \to \infty$. Thus we have

$$0 < -\nabla\left(S_t(\tilde{\sigma})\nabla \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu)\right) < L$$

we have recovered a Lipschitz continuous gradient of a strictly convex function, proposition 4.2.13 will address the issue of having only strict convexity.

The next example will show that the use of a link function when defining the model density can also aide in making assumptions 4.2.4, 4.2.5 hold.

Example 4.2.12.
Let the model density $\tilde{p}_t$ be chosen as generalized student-t and let the parameter of interest be the scale $\tilde{\sigma} > 0$ but now utilize an exponential link function to define the model as follows

$$Y_t = \exp(\tilde{\sigma}_t)X_t$$

where every X_t is independently distributed as a standard t-student distribution with ν degrees of freedom (this is the case of the Beta-t-EGARCH Harvey and Chakravarty (2008)), then the model density would be

$$\tilde{p}_t(y_t|\tilde{\sigma}, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\exp(\tilde{\sigma})} \left(1 + \frac{1}{\nu}\left(\frac{y_t}{\exp(\tilde{\sigma})}\right)^2\right)^{-\frac{\nu+1}{2}}$$

with score given by

$$\nabla \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu) = \frac{(\nu+1)y_t^2}{\nu \exp(2\tilde{\sigma}) + y_t^2} - 1$$

that has second derivative

$$-\nabla^2 \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu) = \frac{(\nu+1)y_t^2 \nu 2 \exp(2\tilde{\sigma})}{(\nu \exp(2\tilde{\sigma}) + y_t^2)^2}$$

and by inspection it is easy to see that

$$0 < -\nabla^2 \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu, \nu) < L$$

for all y_t and all $\tilde{\sigma}$ and where $L \in \mathbb{R}^+$, notice that even here we achieve only strict convexity, the next proposition will address this issue. Notice that also in this case the choice of $S_t := 1$ is in line with the standard practice when defining a Beta-t-EGARCH model Harvey and Chakravarty (2008), section 4.2.2 will elaborate more on this.

Since the t student is often used as a model of choice, especially in financial applications where returns exhibit heavy tails and heteroskedasticity, we wish to extend the theory developed up until now to include even strictly convex objective functions. The following proposition is motivated by this goal.

Proposition 4.2.13.
let assumption 4.2.4, 4.2.2 hold and let

$$0 < -\nabla\left(S_t(\tilde{\lambda}, \boldsymbol{\theta})\nabla \log \tilde{p}(y_t|\tilde{\lambda}, \boldsymbol{\theta})\right)$$

for all $\tilde{\lambda} \in \tilde{\Lambda}, y_t \in \mathbb{R}$ *and all* t*. Then if* $0 < \alpha < 2/L$ *every Newton-score update is CEV optimal for any true density* p_t*, i. e.,*

$$\Delta_{E_{t|t-1}} < 0 \qquad \forall t$$

Consequently even weaker assumptions than 4.2.5 can lead to CEV optimality. This should not surprise us given that the literature on optimization has found convergence properties for the gradient descent scheme outside the simplifying assumptions of strong convexity and a Lipschitz continuous gradient.
The next section aims to generalize the result obtained on CEV optimality to a more general class of score driven models without restricting ourselves only to the Newton-score model 4.2 defined in Blasques et al. (2015).

4.2.2 Adjusted Conditional Expected Variation Optimality

Given that, in practice, observation driven models are often used for prediction, a priority is to investigate under which conditions an update of a score driven model from an arbitrary starting value $\tilde{\lambda}_t$ moves closer, in expectation, to the pseudo-true parameter at the next time λ^*_{t+1}, that is, mirroring the CEV definition, what are the conditions under which $\|\lambda^*_{t+1} - E_{Y_t|Y^{t-1}}\left[\tilde{\lambda}_{t+1}\right]\| < \|\lambda^*_t - \tilde{\lambda}_t\|$. This case was not tackled in Blasques et al. (2015) since it would have required conditions on the evolution of $\{\lambda^*_t\}_{t\in\mathbb{N}}$, in this section we will impose some behavior on the sequence of true parameter values to approach this issue.

Paralleling the assumption in section 2.4.2 but adding a stochastic component we choose the evolution of λ^*_t as that of a stationary AR(1)

Assumption 4.2.14.
Let $\{\lambda^*_t\}_{t\in\mathbb{N}}$ be a stationary autoregressive process of order 1, i. e.,

$$\lambda^*_{t+1} = \omega + \beta\lambda^*_t + \epsilon_t$$

where $\omega \in \mathbb{R}, \beta \in (-1, 1)$ and $\{\epsilon_t\}_{t\in\mathbb{N}}$ is a mean zero white noise process.

The assumption that the pseudo-true time varying parameter follows an autoregressive process is not new in the literature, the linear Gaussian model and the HAR model by Corsi (2009) are two popular examples (although the HAR model is an AR(22)). Moreover, in the case of correct specification, any DGP $\{Y_t\}_{t\in\mathbb{N}}$ that evolves according to the following, commonly used, class of parameter driven models would fall under assumption 4.2.14

$$\begin{aligned} Y_t &= \lambda^*_t + w_t \qquad \forall t \in \mathbb{N} \\ \lambda^*_{t+1} &= \beta\lambda^*_t + \epsilon_t \qquad \forall t \in \mathbb{N} \end{aligned}$$

where $|\beta| < 1$ and the usual assumptions on the sequence of random variables $\{w_t\}_{t\in\mathbb{Z}}, \{\epsilon_t\}_{t\in\mathbb{Z}}$ hold, namely that

$$\forall t_1 \neq t_2 : \ \epsilon_{t_1} \perp \epsilon_{t_2}, w_{t_1} \perp w_{t_2} \forall t_1, t_2 : \ \epsilon_{t_1} \perp w_{t_2}, \forall t : \ E[\epsilon^2_t] = Q, E[w^2_t] = R$$

where the notation $X \perp Y$ signifies that X and Y are independent random variables and Q, R are assumed to be positive real numbers. The same is true for stochastic volatility models of the type

$$\begin{aligned} Y_t &= \lambda^*_t w_t \qquad \forall t \in \mathbb{N} \\ \lambda^*_{t+1} &= \beta\lambda^*_t + \epsilon_t \qquad \forall t \in \mathbb{N} \end{aligned}$$

where $|\beta| < 1$ and the above assumptions on the sequence of random variables $\{w_t\}_{t\in\mathbb{Z}}$, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ hold. Observe that, in both models, w_t and ϵ_t are not necessarily assumed to be normally distributed.

Given the novel auto-regressive assumption 4.2.14 on the behavior of the pseudo-true time varying parameter λ_t^*, we formally define the new property we shall investigate

Definition 4.2.15.
The adjusted conditional expected variation (ACEV) of a parameter update from $\tilde{\lambda}_t \in \tilde{\Lambda} \subseteq R$ to $\tilde{\lambda}_{t+1} \in \tilde{\Lambda} \subseteq R$ at time t is given by

$$\Delta^{adj}_{E_{t|t-1}} = \left\| E_{\epsilon_t}[\lambda^*_{t+1}] - E_{Y_t|Y^{t-1}}\left[\tilde{\lambda}_{t+1}\right] \right\| - \left\| \lambda^*_t - \tilde{\lambda}_t \right\|$$

a parameter update will be said to be ACEV optimal if and only if $\Delta^{adj}_{E_t} < 0$

This definition closely resembles the one in 4.2.1 the biggest difference being that we find another expected value with respect to the random variable ϵ_t that is needed to control the stochastic behavior of the pseudo-true parameter, since we assume it evolves according to equation 4.2.14. If the variance of ϵ_t was zero then we could drop the expectation with respect to ϵ_t altogether, the definition would reduce to the one in 4.2.1 except we have concordance in the times at which the distances between the model time varying parameter and the pseudo true time varying parameter are evaluated. This makes ACEV a more interesting property for predictive purposes than CEV. The following proposition will show that under the same assumptions of section 4.2.1 this property is achieved by specific score driven models of order one, as opposed to before where we proved CEV optimality only for the Newton-score update.

Proposition 4.2.16.
Let assumptions 4.2.4, 4.2.5, 4.2.2, 4.2.14 hold. Then, if $0 < \alpha < (1+\beta)/L$, every update of the first order score driven model given by

$$\tilde{\lambda}_{t+1} = \omega + \beta\tilde{\lambda}_t - \alpha S_t(\tilde{\lambda}, \boldsymbol{\theta})\tilde{s}(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) \tag{4.3}$$

is ACEV optimal for any true density p_t, i. e.,

$$\Delta^{adj}_{E_{t|t-1}} < 0 \qquad \forall t$$

The proof of this proposition utilizes the same techniques of the optimization theory presented in chapter 2, in particular it can be viewed as the stochastic analogue to 2.4.10

in section 2.4.2.

For how theoretically pleasing this proposition might be, in practice, there still remains a glaring flaw: in almost all applied cases we don't know the parameters ω, β that govern the DGP so directly specifying the score driven model as in 4.3 is infeasible. One could think about obtaining estimators for the parameters ω, β under the autoregressive assumption 4.2.14 and substituting them in the score driven specification, essentially utilizing score driven models as one step ahead predictors for parameter driven models. There is already some evidence in the literature supporting this view, see Harvey (2013) where score driven models are cast as filters of unobserved component models and Fleming and Kirby (2003) where the GARCH(1,1) is seen as a filter for a simple stochastic volatility model, we leave this avenue for further study.

Alternatively one can view score driven models as observations driven models that may posses an extra beneficial property (depending on the estimation of the parameters ω, β) even in the case of model misspecification.

As in proposition 4.2.7 S_t must be chosen appropriately for assumptions 4.2.4, 4.2.5 to hold, this diverges from the current literature on score driven models where S_t is usually chosen ad-hoc or following some heuristic guideline. Thus proposition 4.2.7 and proposition 4.2.16 mark the first time, to the authors knowledge, that specific conditions are required on the form of the function S_t for the model specification to meet an optimality property. The next corollary shows how the famous GARCH(1,1) specification is recovered in this framework

Corollary 4.2.17.

Let $\tilde{p}_t(\cdot|\tilde{\sigma}_t^2, \theta))$ be chosen as the mean zero, Gaussian density for all t and assume the pseudo-true time varying parameter of interest $\{(\sigma_t^*)^2\}_{t\in\mathbb{N}}$ evolves according to assumption 4.2.14 then the score driven model

$$\tilde{\sigma}_{t+1}^2 = \omega + \beta\tilde{\sigma}_t^2 - \alpha S_t s_t \tag{4.4}$$

with $0 < \alpha < (1+\beta)/\mathrm{L}$ and $S_t := \tilde{\sigma}_t^4$ (the inverse of the information matrix) has ACEV optimal updates for all t, i.e.,

$$\Delta^{adj}_{E_{t|t-1}} < 0 \qquad \forall t$$

The model in 4.4 has the GARCH(1,1) form since

$$\tilde{\sigma}_{t+1}^2 = \omega + \beta\tilde{\sigma}_t^2 - \alpha S_t s_t = \omega + (\beta+\alpha)\tilde{\sigma}_t^2 - \alpha X_t^2$$

The ARCH case could also be recovered in the case that $\alpha := -\beta$.

As in lemma 2.2.2 we need to modify the gradient descent update to recover assumptions

4.2.4 and 4.2.5 on the sequence of objective functions. This is done through the scaling function S_t, thus S_t must be chosen as to make the gradients of the sequence of objective functions Lipschitz without changing their minimizer. As explained in example 4.2.9 the natural choice is the inverse of the information matrix. Thus corollary 4.2.17 shows that the choice of S_t as the inverse of the information matrix occurs naturally, when requiring ACEV optimality, if the assumption on the distribution of the observations is Gaussianity. We thus recover a novel motivation for the choice of S_t as the inverse of the information matrix, that had already been proposed in the literature Creal et al. (2013), Harvey (2013). To get the point across, we may imagine an econometrician that wishes to utilize score driven models to model heteroskedasticity through time without knowledge of the GARCH specification, if he wanted to achieve ACEV optimality he would choose S_t as to satisfy assumptions 4.2.4, 4.2.5 and the natural choice would then happen to be the inverse of the information matrix, that would then result in a GARCH specification. Thus, in this paradigm, the choice of S_t is not made ad-hoc in order to recover the GARCH specification but it directly follows from the conditions needed to achieve ACEV optimality. Of course it is already well known that the GARCH specification achieves very good results in practice Hansen and Lunde (2005a), consequently the conditions imposed on S_t through assumptions 4.2.4 and 4.2.4 are of greater importance, from an applied perspective, when the distribution of the observations is assumed non Gaussian. Then novel individual specifications of score driven models that achieve ACEV optimality may be found.

As exemplified in 4.2.11 and 4.2.12 sometimes, through the choice of S_t, one only manages to achieve

$$0 < -\nabla \left(S_t(\tilde{\lambda}, \boldsymbol{\theta}) \nabla \log \tilde{p}(y_t | \tilde{\lambda}, \boldsymbol{\theta}) \right) \qquad \forall \tilde{\lambda} \in \tilde{\Lambda}, \forall y_t \in \mathbb{R}, \forall t \in \mathbb{N} \tag{4.5}$$

As was the case for proposition 4.2.7 we can state an analogous proposition to 4.2.16 in this increasingly general case

Proposition 4.2.18.

Let assumptions 4.2.4, 4.2.2, 4.2.14 and equation 4.5 hold then, if $0 < \alpha < (1 + \beta)/L$, *every update of the first order score driven model given by*

$$\tilde{\lambda}_{t+1} = \omega + \beta \tilde{\lambda}_t - \alpha S_t(\tilde{\lambda}, \boldsymbol{\theta}) \tilde{s}(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) \tag{4.6}$$

is ACEV optimal for any true density p_t*, i. e.,*

$$\Delta^{adj}_{E_{t|t-1}} < 0 \qquad \forall t$$

The next corollary shows how the Beta-t-EGARCH model introduced by Harvey Harvey and Chakravarty (2008) is ACEV optimal if the model parameters are chosen correctly.

Corollary 4.2.19.
Let $\tilde{p}_t(\cdot|\tilde{\sigma}_t, \boldsymbol{\theta}))$ be chosen as a generalized t-student utilizing an exponential link function for the scale, i.e.,

$$\tilde{p}_t(y_t|\tilde{\sigma}_t, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\exp(\tilde{\sigma}_t)} \left(1 + \frac{1}{\nu}\left(\frac{y_t}{\exp(\tilde{\sigma}_t)}\right)^2\right)^{-\frac{\nu+1}{2}}, \quad \boldsymbol{\theta} := \nu \tag{4.7}$$

and assume the pseudo-true time varying parameter of interest $\{(\sigma_t^*)\}_{t\in\mathbb{N}}$ evolves according to assumption 4.2.14 then the score driven model

$$\tilde{\sigma}_{t+1} = \omega + \beta\tilde{\sigma}_t - \alpha S_t s_t \tag{4.8}$$

with $0 < \alpha < (1+\beta)/\mathrm{L}$ and $S_t := 1$ has ACEV optimal updates for all t, i.e.,

$$\Delta^{adj}_{E_{t|t-1}} < 0 \qquad \forall t$$

Under the density assumption 4.7 the score of the generalized t student is

$$s_t = \frac{(v+1)Y_t^2}{v\exp(2\tilde{\sigma}_t) + Y_t^2} - 1$$

thus equation 4.8 defines a Beta-t-EGARCH(1,1) model

$$\tilde{\sigma}_{t+1} = \omega + \beta\tilde{\sigma}_t - \alpha S_t s_t = \omega + \beta\tilde{\sigma}_t - \alpha\left(\frac{(v+1)Y_t^2}{v\exp(2\tilde{\sigma}_t) + Y_t^2} - 1\right)$$

Again we remark that the natural choice of S_t to satisfy assumptions 4.2.4, 4.2.5 was the identity (notice that in this case the inverse of the information matrix would not have satisfied assumptions 4.2.4 and 4.2.5) so the corollary provides a novel motivation, analogously to the GARCH case, for the choice of $S_t = 1$ in the specification of the Beta-t-EGARCH(1,1), this is already the standard rule of thumb in the literature Harvey (2013), Harvey and Chakravarty (2008), Creal et al. (2013).

4.2.3 The Invertibility Conjecture

Let assumption 4.2.14 hold with zero variance of the error term ϵ_t and assume a model of interest is ACEV optimal, i.e., we have that

$$\|\lambda^*_{t+1} - E_{Y_t|Y^{t-1}}[\tilde{\lambda}_{t+1}]\| < \|\lambda^*_t - \tilde{\lambda}_t\|$$

ACEV optimality in this case tells us that at every step t the model predicted parameter $\tilde{\lambda}_{t+1}$ gets closer than the previous one $\tilde{\lambda}_t$ to the time varying pseudo-true parameter value. One could then expect that

$$\|\tilde{\lambda}_t - \lambda^*_t\| \xrightarrow{\mathbb{P}} 0$$

This would mean that an ACEV optimal model is invertible, to prove this kind of result we conjecture that it is possible to adjust the proofs 2.3.2, 2.3.4 utilized to prove convergence of the stochastic gradient descent scheme in a time varying objective function setting, we leave this avenue for further research.

Notice thought that in this setting we are making an assumption on the behavior of λ^*_t through time, this is unlike the canonical situation that has been analyzed when proving invertibility of observation driven models Strauman and Mikosch (2006), Wintenberger (2013), Blasques et al. (2016a) where no assumption on the behavior of λ^*_t is made.

Chapter 5

Concluding Thoughts on Score Driven Models and Further Research

The use of the derivative of the log-likelihood with respect to the parameter of interest in the updating equation of observation driven models has found success in applications. On the other hand, the theoretical implications of the use of the score in the dynamic equation are yet to be completely uncovered. In this thesis we have analyzed the updating equation of the score driven class of models and derived a global, high dimensional property arising by the use of score in the updating equation.

To formalize CEV and ACEV optimality we established a connection with the optimization literature. In fact, choosing a model to track a time varying parameter and selecting a time varying stochastic optimization scheme to optimize the sequence of expected log-likelihoods are two problems, that under the appropriate conditions, have a related goal.

To conclude the thesis, some remarks including further avenues of research can be spelled out.

1. The time varying optimization problem defined on the sequence of one observation log-likelihoods is a generalization of maximum likelihood theory, where a fixed parameter is being estimated, to the case where the parameter to be estimated varies through time.

2. CEV optimality leads to naturally interpret score driven models as filters for the time varying parameter of interest, that we can assume to evolve according to an unknown data generating process. Conversely ACEV optimality characterizes

score driven models as one step ahead predictors by assuming that the quantities that govern the evolution of the true time varying parameter are known. In this sense, score driven models appear to act naturally as filters or predictors for state space models.

3. ACEV optimality also shows how score driven models may be interpreted as a prediction correction scheme, like the one analyzed in section 2.4.1. The autoregressive component being the predictive step that approximates the evolution of the time varying maxima. While the score component is pushing the dynamics in the direction of steepest ascent.

$$\tilde{\lambda}_{t+1} = \underbrace{\gamma + \sum_{i=0}^{q} \beta_i \tilde{\lambda}_{t-i}}_{\text{prediction}} + \underbrace{\sum_{i=0}^{p} \alpha_i S_{t-i} \frac{\partial \log \tilde{p}(Y_{t-i}|\tilde{\lambda}_{t-i};\theta)}{\partial \tilde{\lambda}_{t-i}}}_{\text{correction}}$$

4. A further avenue of research to be investigated when using the Newton-score update 4.2, since it's CEV optimal for every t, is to perform multiple updates in the same time frame, i. e., for a fixed t one could execute $\tau \in \mathbb{N}$ Newton-score updates (gradient descent updates)

$$\tilde{\lambda}_{t_{h+1}} = \tilde{\lambda}_{t_h} + \alpha S_t(\tilde{\lambda}_{t_h}, \boldsymbol{\theta}) s_t(Y_t, \tilde{\lambda}_{t_h}, \boldsymbol{\theta})$$

where τ is defined on the basis of a stopping rule. This procedure would be analogous to what is canonically done in the optimization literature, then one could set $\tilde{\lambda}_{t_{\tau+1}} := \tilde{\lambda}_{t+1}$. Generalizing, in this way, the Newton-score update. The same reasoning could then be extended to any arbitraty CEV optimal model.

5. A further avenue of research spurred by the optimization theory connection is that different observation driven models could be inspired by optimization schemes and retain desirable properties, just as CEV optimality is retained as a consequence of the monotonicity of the gradient descent updates. For example the following specification given in the 1-dimensional setting mirrors Newton's method

$$\tilde{\lambda}_{t+1} = \gamma + \sum_{i=0}^{p} \alpha_i S_{t-i} \frac{\partial \log \tilde{p}(Y_{t-i}|\tilde{\lambda}_{t-i}, \theta)}{\partial \tilde{\lambda}_{t-i}} \frac{\partial^2}{\partial \log \tilde{p}(Y_{t-i}|\tilde{\lambda}_{t-i}, \theta)} + \sum_{i=0}^{q} \beta_i \tilde{\lambda}_{t-i}$$

Interestingly, it is well known that Newtons method converges faster than gradient descent, with the caveat of needing the second derivative of the objective function, and the scheme has been already generalize to the time varying parameter setting Simonetto et al. (2016), so an observation driven model for a time varying parameter based on Newtons method could have beneficial theoretical properties.

All these points are left for further research and discussion.

Appendix A

Most proofs in this appendix are standard, when non-standard or original proofs will be given we will explicitly remark it.

Proof of lemma 2.1.1

We assume first that f is Lipschitz, through an application of the fundamental theorem of calculus we have

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \frac{\partial f(x - t(-y+x))}{\partial t}\, dt \\
&= f(x) + \int_0^1 \nabla f(x - t(y-x))(y-x)\, dt \\
&= f(x) + \nabla f(x)(y-x) + \int_0^1 (\nabla f(x - t(y-x)) - \nabla f(x))(y-x)\, dt \\
&\leq f(x) + \nabla f(x)(y-x) + \int_0^1 \|\nabla f(x - t(y-x)) - \nabla f(x)\| \|y-x\|\, dt \\
&\leq f(x) + \nabla f(x)(y-x) + \int_0^1 \|y-x\| tL \|y-x\|\, dt \\
&\leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}\|x-y\|^2
\end{aligned}$$

where the Lipschitz assumption was used in the second to last inequality. Notice that for this direction the convexity was not needed.

Now for the converse, define a function $\phi(y) := f(y) - y\nabla f(x)$ that is convex since its an affine function of $f(y)$. Notice that this function achieves a global minimum at x since $\nabla\phi(x) = \nabla f(x) - \nabla f(x) = 0$ in addition

$$\phi(y) \leq \phi(z) + \nabla\phi(z)(y-z) + \frac{L}{2}\|y-z\|^2 \tag{A.1}$$

since

$$f(y) - y\nabla f(x) \leq f(z) - z\nabla f(x) + (\nabla f(z) - \nabla f(x))(y-z) + \frac{L}{2}\|y-z\|^2$$

follows by assumption. Since x is a global minimum we have that

$$\begin{aligned}&\phi(x) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \\ &\phi(y) + \nabla\phi(y)\left(y - \frac{1}{L} - y\right)\nabla\phi(y) + \frac{1}{2L}\|\nabla\phi(y)\|^2 = \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|^2\end{aligned}$$

where we have used equation A.1 in the second inequality, substituting the definition of ϕ we obtain

$$f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|$$

summing this inequality with the same inequality with x and y reversed we obtain

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq (\nabla f(x) - \nabla f(y))(x-y)$$

and by applying Cauchy Schwarz we have

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq (\nabla f(x) - \nabla f(y))(x-y) \leq \|\nabla f(x) - \nabla f(y)\|\|x-y\|$$

that gives us the Lipschitz condition.

Proof of lemma 2.1.2

We prove only the bound equivalent to strong convexity since the argument for the bound given by a Lipschitz continuous gradient is analogous given the symmetry in the equivalent definition given by lemma 2.1.1. We notice that the condition

$$f(x) \geq f(y) + \nabla f(y)(x-y) + \frac{\ell}{2}||x-y||^2 \quad \forall x, y$$

is equivalent to the function $g(x) := f(x) - \frac{\ell}{2}\|x\|^2$ being convex since

$$\begin{aligned}&g(x) - g(y) \geq \nabla g(y)(x-y) \iff \\ &f(x) - \frac{\ell}{2}\|x\|^2 - f(y) + \frac{\ell}{2}\|y\|^2 \geq (\nabla f(y) - \ell y)(x-y) \iff \\ &f(x) \geq f(y) + \nabla f(y)(x-y) + \frac{\ell}{2}||x-y||^2\end{aligned}$$

now define $h(y) := g(y) - \nabla g(x)(y-x)$ this is another convex function since it's an affine mapping of the convex function g, in addition

$$\nabla h(y) = \nabla g(y) - \nabla g(x), \quad \nabla^2 h(y) = \nabla^2 g(y)$$

In particular $\nabla h(x) = 0$ thus x is a global minimum of h, from the necessary conditions of minimum points this implies that $\nabla^2 h(x)$ is positive semi-definite but $\nabla^2 h(x) = \nabla^2 g(x)$

so even $\nabla^2 g(x)$ is positive semi-definite and this implies that $\nabla^2 f - \ell I_{d\times d} \geq 0$.
Now we have to show the converse, i. e., if $\nabla^2 f(x) \geq \ell I_{d\times d}$ then f is strongly convex, through an application of Taylor's theorem we have

$$f(x) = f(y) + \nabla f(y)(x-y) + \frac{1}{2}(x-y)\nabla^2 f(y + t(x-y)(x-y) \pm \frac{1}{2}(x-y)I_{d\times d}\ell(x-y)$$

for some $0 \leq t \leq 1$. Since $(\nabla^2 f(y + t(x-y) - I_{d\times d}\ell)$ is positive semi-definite we obtain the bound

$$f(x) \geq f(y) + \nabla f(y)(x-y) + \frac{\ell}{2}\|x-y\|^2$$

this ends the proof.

Proof of lemma 2.1.3
Through an application of the fundamental theorem of calculus

$$\begin{aligned}\nabla f(x) &= \nabla f(y) + \int_0^1 \frac{\partial \nabla f(y - t(x-y))}{\partial t}\, dt \\ &= \nabla f(y) + \int_0^1 \nabla^2 f(y - t(x-y))^T (x-y)\, dt \\ &= \nabla f(y) + A(x-y)\end{aligned}$$

where $A = \int_0^1 \nabla^2 f(y - t(x-y))^T\, dt$ is $I_{d\times d}\ell \leq A \leq I_{d\times d}\mathrm{L}$ from the bounds on the Hessian function.

Proof of lemma 2.1.4
From the strong convexity we have

$$\begin{aligned}f(x^*) - f(x) &\geq \nabla f(x)(x^* - x) + \frac{\ell}{2}||x^* - x||^2 \\ f(x) - f(x^*) &\geq \nabla f(x^*)(x - x^*) + \frac{\ell}{2}||x^* - x||^2\end{aligned}$$

summing both inequalities gives

$$(\nabla f(x) - \nabla f(x^*))(x - x^*) = \nabla f(x)(x - x^*) \geq \ell ||x - x^*||.$$

Proof of lemma 2.1.5
For the left hand side inequality we minimize both sides of the inequality in 2.0.2 with respect to x yielding

$$f(x^*) \geq f(x) - \frac{1}{2\ell}\left(\nabla f(x)\right)^2$$

that re-arranged gives the PL inequality.
For the right hand side we notice that substituting the minima $y = x^*$ in assumption 2.0.2 immediately implies that

$$f(x) \geq f(x^*) + \ell(x - x^*)^2/2$$

since $\nabla f(x^*) = 0$.

Proof of lemma 2.1.6
As in the previous proof of lemma 2.1.5: the right had side inequality follows immediately choosing $y = x^*$ in the inequality in 2.0.1.
The left had side inequality follows by minimizing both sides of the inequality in 2.0.1 with respect too x

$$\begin{aligned}&\min_x f(x) \leq \min_x \left(f(y) + \nabla f(y)(y-x) + \frac{1}{2L}||y-x||^2 \right) \implies \\ &f(x^*) \leq f(x) - \frac{1}{2L}||\nabla f(x)||^2\end{aligned}$$

Proof of lemma 2.1.7
the proof follows the same argument as lemma 2.1.1, define two convex functions

$$f_x(z) := f(z) - \nabla f(x)^T z, \quad f_y(z) := f(z) - \nabla f(y)^T z$$

these functions are convex because

$$\begin{aligned}&f_x(\alpha z_1 + \beta z_2) := f(\alpha z_1 + \beta z_2) - \nabla f(x)^T(\alpha z_1 + \beta z_2) \leq \\ &f(\alpha z_1) + f(\beta z_2) - \nabla f(x)^T \alpha z_1 - \nabla f(x)^T \beta z_2 = f_x(\alpha z_1) + f_x(\beta z_2)\end{aligned}$$

for any $\alpha, \beta \in \mathbb{R}^+$ such that $\alpha + \beta = 1$. Furthermore it is immediate that the two functions have Lyschitz continuous gradient, so

$$\begin{aligned}&f(y) - f(x) - \nabla f(x)^T(x-y) = f_x(y) - f_x(x) \geq \\ &\frac{1}{2L}||\nabla f_x(y)||^2 = \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2\end{aligned}$$

similarly we can find

$$f(x) - f(y) - \nabla f(y)^T(y-x) \geq \frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2$$

summing these last two inequalities we obtain the result.

Proof of proposition 2.2.1

Utilizing lemma 2.1.3 and choosing $y = x^*, x = x_k$ one obtains

$$\nabla f(x_k) = \nabla f(x^*) + A_k(x_k - x^*) = A_k(x_k - x^*).$$

Recalling that the gradient descent update is $x_{k+1} = x_k - \alpha\nabla f(x_k)$, we have

$$||x_{k+1} - x^*||_2 = ||x_k - \alpha\nabla f(x_k) - x^*||_2 = ||(I_{d\times d} - \alpha A_k)(x_k - x^*)||_2 \leq$$
$$||I_{d\times d} - \alpha A_k||||x_k - x^*||_2.$$

For every symmetric matrix A it is true that $||I - A|| = \max\{|1-\lambda_1|, |1-\lambda_n|\}$ where λ_1, λ_n are respectively the smallest and largest eigenvalues of A. Hence $||x_{k+1} - x^*||_2 \leq q||x_k - x^*||_2, q = \max\{|1-\lambda_1|, |1-\lambda_n|\}$. Since $0 < \alpha < 2/\mathrm{L}$ and $0 < \ell < \mathrm{L}$ then $|1 - \alpha\ell| < 1, |1 - \alpha\mathrm{L}| < 1$ so $q < 1$.
Minimizing q over α we obtain $q = \max\{|1-\alpha\ell|, |1-\alpha\mathrm{L}|\} < 1$.

Proof of lemma 2.2.2

This is an original modification of 2.2.1.
As in the proof of lemma 2.1.3

$$\begin{aligned} g(\nabla f)(x) &= g(\nabla f)(y) + \int_0^1 \frac{\partial g(\nabla f)(y - t(x-y))}{\partial t}\, dt \\ &= g(\nabla f)(y) + \int_0^1 \nabla(g(\nabla f)(y - t(x-y)))^T(x-y)\, dt \\ &= g(\nabla f)(y) + A(x-y) \end{aligned}$$

where $A := \int_0^1 \nabla(g(\nabla f)(y - t(x-y)))^T\, dt$ is such that $I_{d\times d}\ell \leq A \leq I_{d\times d}\mathrm{L}$ from the hypothesis.
The rest of the proof mirrors the previous proof of proposition 2.2.1 since

$$||x_{k+1} - x^*||_2 = ||x_k - \alpha g(\nabla f)(x_k) + \alpha g(\nabla f)(x^*) - x^*||_2 = ||(I_{d\times d} - \alpha A_k)(x_k - x^*)||_2 \leq$$
$$||I_{d\times d} - \alpha A_k||||x_k - x^*||_2$$

where we have used that $\alpha g(\nabla f)(x^*) = 0$.

Proof of proposition 2.3.2

Take the squared norm of the difference from the $t+1$ step in the scheme and the minimum x^*

$$||x_{t+1} - x^*||^2 = ||x_t - \alpha_t g(X_t, x_t) - x^*||^2 =$$
$$||x_t - x^*||^2 - 2\alpha_t g(X_t, x_t)(x_t - x^*) + \alpha_t^2||g(X_t, x_t)||^2.$$

Taking conditional expectations produces

$$E_{X_t|X^{t-1}}[||x_t - x^*||^2] - 2\alpha_t \nabla f(x_t)(x_t - x^*) + \alpha_t^2 E_{X_t|X^{t-1}}[||g(X_t, x_t)||^2] \tag{A.2}$$

from the strong convexity assumption 2.0.2 we obtain the bound

$$-\nabla f(x_t)(x_t - x^*) \leq f(x^*) - f(x_t) - \frac{\ell}{2}\|x^* - x_t\|^2$$

utilizing this bound in A.2 we recover

$$E_{X_t|X^{t-1}}[||x_{t+1} - x^*||^2] \leq (1 - \alpha_t \ell)||x_t - x^*||^2 - 2\alpha_t(f(x_t) - f(x^*)) + \alpha_t^2 E_{X_t|X^{t-1}}[||g(X_t, x_t)||^2]$$

taking total expectations and substituting in assumption 2.3.1 we have

$$\begin{aligned} &E[||x_{t+1} - x^*||^2] \leq (1 - \alpha_t \ell)E[||x_t - x^*||^2] + \alpha_t(\alpha_t M_0 - 2)E[(f(x_t) - f(x^*))] + \alpha_t^2 M \\ &\leq (1 - \alpha_t \ell)E[||x_t - x^*||^2] + \alpha_t^2 M \end{aligned}$$

where the last inequality follows since $\alpha_t \leq 2/M_0$ for all t by assumption. Now proceeding by induction we have for the base case that $E[||x_1 - x^*||^2] = ||x_1 - x^*||^2 \leq v$ by the definition of v and the fact that the initial point $x_1 \in \mathbb{R}^d$ is deterministic. Assuming that $E[||x_t - x^*||^2] \leq v/t$ we prove the inductive case

$$\begin{aligned} E[||x_{t+1} - x^*||^2] &\leq \left(1 - \frac{c\ell}{t}\right) E[||x_t - x^*||^2] + \frac{c^2}{t^2}M \\ &\leq \left(1 - \frac{c\ell}{t}\right) \frac{v}{t} + \frac{c^2}{t^2}M \end{aligned}$$

since $c < 1/\ell$ and we have used the inductive assumption. Then

$$\left(1 - \frac{c\ell}{t}\right) \frac{v}{t} + \frac{c^2}{t^2}M = \left(\frac{t-1}{t^2}\right) v - \left(\frac{c\ell - 1}{t^2}\right) v + \frac{c^2}{t^2}M \leq \left(\frac{t-1}{t^2}\right) v$$

where the last inequality follows from the fact that $v \geq c^2M/(c\ell - 1)$. Now noticing that $t^2 \geq (t+1)(t-1)$ we obtain that

$$E[||x_{t+1} - x^*||^2] \leq \left(\frac{t-1}{t^2}\right) v \leq \frac{v}{t+1}$$

this concludes the proof.

Proof of proposition 2.3.4

Define $h_t := ||x_{t+1} - x^*||^2$, we have

$$h_{t+1} - h_t = \alpha_t^2 \|g(X_t, x_t)\|^2 - 2\alpha_t(x_t - x^*)g(X_t, x_t) \tag{A.3}$$

taking conditional expectations

$$E_{X_t|X^{t-1}}[h_{t+1}-h_t] = \alpha_t^2 E_{X_t|X^{t-1}}[\|g(X_t,x_t)\|^2] - 2\alpha_t(x_t-x^*)\nabla f(x_t) \leq \\ \alpha_t^2(M+M_1||\nabla f(x_t)||^2) - 2\alpha_t(x_t-x^*)\nabla f(x_t) \tag{A.4}$$

where assumption 2.3.1 has been used. Applying lemma 2.1.4 and lemma 2.1.6 to A.4 we obtain

$$E_{X_t|X^{t-1}}[h_{t+1}-(1+\alpha_t^2 M_1 L^2)h_t] \leq \alpha_t^2 M - 2\alpha_t \ell ||x_t-x^*||^2 \leq \alpha_t^2 M \tag{A.5}$$

at this point we notice that $E[h_t]$ is bounded, a fact that we will use later, indeed by taking total expectations and recursively substituting we have

$$E[h_{t+1}] \leq E[h_t](1+\alpha_t^2 M_1 L^2) + \alpha_t^2 M \leq \\ E[h_1]\prod_{i=1}^{t}(1+\alpha_i^2 M_1 L^2) + \sum_{i=1}^{t-1}\alpha_i^2 M \prod_{j=i+1}^{t}(1+\alpha_j^2 M_1 L^2) + \alpha_t^2 M$$

and this quantity converges since $E[h_1] = h_1$ is a constant (we start the algorithm at a chosen point), $\prod_{i=1}^{t}(1+\alpha_i^2 M_1 L^2)$ converges given that

$$\log\left(\prod_{i=1}^{t}(1+\alpha_i^2 M_1 L^2)\right) = \sum_{i=1}^{t}\log(1+\alpha_i^2 M_1 L^2) \leq \sum_{i=1}^{t}\alpha_i^2 M_1 L^2 \tag{A.6}$$

and $\sum_{i=1}^{t}\alpha_i^2 M_1 L^2$ converges by assumption, thus also $\sum_{i=1}^{t-1}\alpha_i^2 M\prod_{j=i+1}^{t}(1+\alpha_j^2 M_1 L^2)$ converges as $\prod_{j=i+1}^{t}(1+\alpha_j^2 M_1 L^2)$ is bounded and the last term $\alpha_t^2 M$ trivially goes to zero. Now define

$$u_t := \prod_{i=1}^{t-1}\frac{1}{1+\alpha_i^2 M_1 L^2}, \qquad h_t' = u_t h_t$$

notice that u_t converges to a strictly positive quantity since following a similar argument as before

$$\log(u_t) = -\sum_{i=1}^{t-1}\log(1+\alpha_i^2 M_1 L^2) \geq -\sum_{i=1}^{t-1}\alpha_i^2 M_1 L^2$$

and the right hand side converges since $\sum_{i=1}^{\infty}\alpha_i < \infty$ implying that u_t converges to a strictly positive quantity $u_\infty > 0$. Multiplying both sides of A.5 by u_t we obtain

$$E_{X_t|X^{t-1}}[h_{t+1}' - h_t'] \leq \alpha_t^2 u_t M$$

from which we have that

$$E[\mathbb{1}_{F_t}(h_{t+1}' - h_t')] = E[\mathbb{1}_{F_t} E_{X_t|X^{t-1}}[h_{t+1}' - h_t']] \leq \alpha_t^2 u_t M$$

where $F_t := \{\omega \in \Omega | E_{X_t|X^{t-1}}[h'_{t+1} - h'_t] > 0\}$. So

$$\sum_{i=1}^{\infty} E[\mathbb{1}_{F_i}(h'_{i+1} - h'_t)] \leq \sum_{i=1}^{\infty} \alpha_i^2 u_i M < \infty$$

again by the fact that $\sum_{i=1}^{\infty} \alpha_i < \infty$ and $u_t \to u_\infty$, thus theorem 2.3.3 implies h'_t converges almost surely. From this h_t also converges almost surely since $u_t \to u_\infty > 0$ (here the positiveness is important).

At this point we just need to show that h_t converges to zero, to do this we apply conditional expectation plus assumption 2.3.1 and subsequently total expectations to equation A.3 obtaining

$$E[h_{t+1}] \leq E[h_t] + \alpha_t^2 E[h_t] M_1 L^2 - 2\alpha_t E[(x_t - x^*)\nabla f(x_t)] + \alpha_t^2 M$$

that through recursive substitution becomes

$$E[h_{t+1}] \leq E[h_1] + \sum_{i=1}^{t} \alpha_i^2 E[h_i] M_1 L^2 - 2\sum_{i=1}^{t} \alpha_i E[(x_i - x^*)\nabla f(x_i)] + \sum_{i=1}^{t} \alpha_i^2 M$$

the boundedness of $E[h_t]$ proven at equation A.6 gives us convergence of the series $\sum_{i=1}^{t} \alpha_i^2 E[h_i] M_1 L^2$ while $\sum_{i=1}^{t} \alpha_i^2 M$ converges by assumption this implies that

$$\sum_{i=1}^{\infty} \alpha_i E[(x_i - x^*)\nabla f(x_i)] < \infty$$

and by the assumption that $\sum_{i=1}^{\infty} \alpha_i = \infty$ it must be true that $E[(x_i - x^*)\nabla f(x_i)] \to 0$, it is well known in probability theory that we can extract from a sequence convergent in mean a subsequence x_{i_t} converging almost surely, so $(x_{i_t} - x^*)\nabla f(x_{i_t}) \to 0$ but

$$(x_{i_t} - x^*)\nabla f(x_{i_t}) \geq \ell h_{i_t} \geq 0$$

this shows that $h_{i_t} \to 0$ thus $h_t \to 0$ almost surely concluding the proof.

Proof of lemma 2.4.1

Without loss of generality we can consider the case where

$$u_{t+1} \leq q_0 u_t + q_1 u_{t-1} + \cdots + q_p u_{t-p}$$

since setting $u^* = \epsilon/(1 - q_0 - q_1 - \cdots - q_p)$ one obtains that

$$(u_{t+1} - u^*) \leq q_0(u_t - u^*) + q_1(u_{t-1} - u^*) + \cdots + q_p(u_{t-p} - u^*)$$

It is well known in the recurrent sequence literature that a sequence $\{u_t\}_{t\in\mathbb{N}}$ is a solution to the recurrence relation

$$u_{t+1} = q_0 u_t + q_1 u_{t-1} + \cdots + q_p u_{t-p}$$

if and only if

$$u_{t+1} = a_0 r_0^{t+1} + a_1 r_1^{t+1} + \cdots + a_p r_p^{t+1}$$

where $a_0, a_1, \ldots, a_p$ are real numbers and $r_0, r_1, \ldots, r_p$ are roots of the characteristic equation

$$r^{p+1} - q_0 r^p - q_1 r^{p-1} - \cdots - q_p = 0$$

thus

$$u_{t+1} \leq a_0 r_0^{t+1} + a_1 r_1^{t+1} + \cdots + a_p r_p^{t+1}$$

and the right hand side goes to zero as t goes to infinity since all the roots are assumed by hypothesis inside the unit circle.

Proof of proposition 2.4.2

As in the proof of proposition 2.2.1 we recover that

$$||x_{t+1} - x_t^*|| = ||x_t - \alpha \nabla f_t(x_t) - x_t^*|| \leq q||x_t - x_t^*||$$

this yields

$$||x_{t+1} - x_{t+1}^*|| \leq ||x_{t+1} - x_t^*|| + ||x_t^* - x_{t+1}^*|| \leq q||x_t - x_t^*|| + a$$

utilizing lemma 2.4.1 setting $u_t := ||x_t - x_t^*||$ we obtain the result.

Proof of proposition 2.4.3

As in the proof of proposition 2.4.2 we easily recover

$$||x_{t+1} - x_t^*|| = ||x_t - \alpha \nabla f_t(x_t) - x_t^*|| \leq q||x_t - x_t^*||$$

and we have

$$||x_{t+1} - x_{t+1}^*|| \leq ||x_{t+1} - x_t^*|| + ||x_t^* - x_{t+1}^*|| \leq q||x_t - x_t^*|| + ||x_t^* - x_{t+1}^*|| \tag{A.7}$$

from lemma 2.1.5 we know that $||\nabla f_t(x_t)|| \geq \ell \|x_t - x_t^*\|$ from which we have that

$$\|\nabla f_{t+1}(x_{t+1}^*) - \nabla f_t(x_{t+1}^*)\| \geq \ell \|x_{t+1}^* - x_t^*\|$$

since $\nabla f_{t+1}(x_{t+1}^*) = 0$. Consequently from equation A.7 we have

$$||x_{t+1} - x_{t+1}^*|| \leq q||x_t - x_t^*|| + \frac{1}{\ell}\|\nabla f_{t+1}(x_{t+1}^*) - \nabla f_t(x_{t+1}^*)\| = q||x_t - x_t^*|| + \frac{h}{\ell}$$

now through lemma 2.4.1 we have the result.

Proof of lemma 2.4.7

Let us begin by analyzing the forward method Euler method applied to a general vector-valued dynamical system

$$\dot{x} = F(x(t), t) \tag{A.8}$$

if we apply the forward Euler method starting at a certain point $x(t_k)$ we obtain

$$x_{k|k+1} = x(t_k) + hF(x(t_k), t_k) \tag{A.9}$$

on the other hand, we can write $x(t_{k+1})$ through a Taylor expansion as

$$x(t_{k+1}) = x(t_k) + hF(x(t_k), t_k) + \frac{h^2}{2}\frac{d}{dt}F(x(t), t)|_{t=s} \tag{A.10}$$

for a certain time $s \in [t_k, t_{k+1}]$ since $F(x(t_k), t_k)$ is the first derivative of $x(t_k)$.
Subtracting the equality A.9 from A.10 and taking the norm implies that

$$||x_{k|k+1} - x(t_{k+1})|| = ||\Delta_k|| = \left|\left|\frac{h^2}{2}\frac{d}{dt}F(x(t), t)|_{t=s}\right|\right|$$

From the chain rule for multivariate functions we know that

$$\frac{d}{dt}F(x(t), t) = \nabla_t F(x, t) + \nabla_x F(x, t)\dot{x} = \nabla_t F(x, t) + \nabla_x F(x, t)F(x(t), t)$$

where in the last equality we have used equation A.8.
An application of the triangle inequality yields

$$\left|\left|\frac{d}{dt}F(x(t), t)\right|\right| \leq ||\nabla_t F(x, t)|| + ||\nabla_x F(x, t)F(x(t), t)|| \tag{A.11}$$

our goal will be to upper bound the right hand side, to do this we recall that our specific continuous dynamical system in A.8 is given by

$$F(x(t), t) = -\left[\nabla_{xx} f_t(x)\right]^{-1} \nabla_{tx} f_t(x)$$

applying the chain rule to the partial derivative with respect to time results in

$$\nabla_t F(x, t) = -\nabla_t \left[\left[\nabla_{xx} f_t(x)\right]^{-1} \nabla_{tx} f_t(x)\right] =$$
$$\left[\nabla_{xx} f_t(x)\right]^{-1} \nabla_{ttx} f_t(x) - \left[\nabla_{xx} f_t(x)\right]^{-2} \nabla_{txx} f_t(x) \nabla_{tx} f_t(x)$$

taking norms and the triangle inequality

$$||\nabla_t F(x,t)|| \leq || \left[\nabla_{xx} f_t(x)\right]^{-1} \nabla_{ttx} f_t(x)|| + || \left[\nabla_{xx} f_t(x)\right]^{-2} \nabla_{txx} f_t(x) \nabla_{tx} f_t(x)||$$

now utilizing both the bounds present in assumptions 2.4.4 and 2.4.5 we are left with

$$||\nabla_t F(x,t)|| \leq \frac{C_0 C_2}{\ell^2} + \frac{C_3}{\ell} \tag{A.12}$$

applying the same line of reasoning to the second component of the right-hand side of A.11 yields

$$\begin{aligned}&||\nabla_x F(x,t) F(x(t),t)|| = ||(\left[\nabla_{xx} f_t(x)\right]^{-1} \nabla_{xtx} f_t(x) - \left[\nabla_{xx} f_t(x)\right]^{-2} \nabla_{xxx} f_t(x) \nabla_{tx} f_t(x)) F(x(t),t)|| = \\ &|| - \left[\nabla_{xx} f_t(x)\right]^{-2} \nabla_{tx} f_t(x) \nabla_{xtx} f_t(x) + \left[\nabla_{xx} f_t(x)\right]^{-3} \nabla_{xxx} f_t(x) \left[\nabla_{tx} f_t(x)\right]^2 || \leq \\ &\frac{C_0 C_2}{\ell^2} + \frac{C_0^2 C_1}{\ell^3}\end{aligned} \tag{A.13}$$

where in the last equality we have again used assumptions 2.4.4 and 2.4.5. At this point the result holds by combining the upper bounds of A.12 and A.13.

Proof of proposition 2.4.8

For the sake of simplicity define

$$\begin{aligned}\nabla_{xx} f &:= \nabla_{xx} f_{t_k}(x_k), & \nabla_{tx} f &:= \nabla_{tx} f_{t_k}(x_k), \\ \nabla_{xx} f^* &:= \nabla_{xx} f_{t_k}(x^*(t_k)), & \nabla_{xx} f^* &:= \nabla_{tx} f_{t_k}(x^*(t_k))\end{aligned}$$

where $x^*(t_k)$ is the minimum of the function f_{t_k}.
Let's begin by evaluating the error term coming from the approximate time derivative in 2.10. In particular consider the Taylor expansion of the gradient $\nabla_x f_{t_{k-1}}(x_k)$ with Lagrange remainder.

$$\nabla_x f_{t_{k-1}}(x_k) = \nabla_x f_{t_k}(x_k) - h \nabla_{tx} f_{t_k}(x_k) + \frac{h^2}{2} \nabla_{ttx} f_s(x_k)$$

where $s \in [t_{k-1}, t_k]$. Rearranging we see that the partial mixed gradient can be written as

$$\nabla_{tx} f_{t_k}(x_k) = \frac{\nabla_x f_{t_k}(x_k) - \nabla_x f_{t_{k-1}}(x_k)}{h} + \frac{h}{2} \nabla_{ttx} f_s(x_k)$$

from the definition of approximate partial mixed gradient this becomes

$$\nabla_{tx} f_{t_k}(x_k) - \tilde{\nabla}_{tx} f_{t_k}(x_k) = \frac{h}{2} \nabla_{ttx} f_s(x_k)$$

we know from assumption 2.4.5 that $\nabla_{ttx} f_s(x_k)$ is bounded above by C_3 therefore

$$||\nabla_{tx} f_{t_k}(x_k) - \tilde{\nabla}_{tx} f_{t_k}(x_k)|| \leq \frac{hC_3}{2} \tag{A.14}$$

we have thus successfully bounded the error term of the approximate time derivative 2.10.

Then adding and subtracting the exact prediction direction $h\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f$ from the definition of $x_{k+1|k}$ in 2.6 we have

$$x_{k+1|k} = x_k - h\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f + h\left[\nabla_{xx} f\right]^{-1} \left(\nabla_{tx} f - \tilde{\nabla}_{tx} f\right)$$

at this point we subtract

$$x^*_{k+1|k} + x^*(t_{k+1}) = x^*(t_k) - h\left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f^* + x^*(t_{k+1})$$

that is the definition of $x_{k+1|k}$ knowing $\nabla_{tx} f^*$ to which we have added on both sides $x^*(t_{k+1})$, the asterisk notation $x^*_{k+1|k}$ indicates that we start from the minimum $x^*(t_k)$. Computing the subtraction we obtain

$$\begin{aligned} &x_{k+1|k} - x^*(t_{k+1}) = x_k - x^*(t_k) + \\ &h\left(\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f^*\right) + h\left[\nabla_{xx} f\right]^{-1} \left(\nabla_{tx} f - \tilde{\nabla}_{tx} f\right) + x^*_{k+1|k} - x^*(t_{k+1}) \end{aligned}$$

considering the norm and utilizing the triangle inequality leads to

$$\begin{aligned} &||x_{k+1|k} - x^*(t_{k+1})|| \leq ||x_k - x^*(t_k)|| + \\ &h||\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f^*|| + h||\left[\nabla_{xx} f\right]^{-1} (\nabla_{tx} f - \tilde{\nabla}_{tx} f)|| + ||\Delta^*_k|| \end{aligned} \tag{A.15}$$

where $\Delta^*_k = x^*_{k+1|k} - x^*(t_{k+1})$. We can upper bound the error norm $||\Delta^*_k||$ through lemma 2.4.7 also from A.14 and assumption 2.4.5 we can bound the third term

$$||\left[\nabla_{xx} f\right]^{-1} (\nabla_{tx} f - \tilde{\nabla}_{tx} f)|| \leq \frac{C_3 h}{2\ell}$$

now we turn our attention to the final term $||\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f^*||$. First we add and subtract $\left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f$ then we use the triangle inequality to obtain

$$\begin{aligned} &||\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f^*|| \leq \\ &||\left[\nabla_{xx} f\right]^{-1} \nabla_{tx} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f|| + ||\left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{tx} f^*|| \leq \\ &C_0||\left[\nabla_{xx} f\right]^{-1} - \left[\nabla_{xx} f^*\right]^{-1}|| + \frac{1}{\ell}||\nabla_{tx} f - \nabla_{tx} f^*|| \end{aligned}$$

where in the last inequality we have used both assumption 2.4.4 and 2.4.5.
We now further bound the first term of the right-hand side. To do that, we use the non-singularity of the Hessian to write

$$||\,[\nabla_{xx}f]^{-1}-[\nabla_{xx}f^*]^{-1}\,||=||\,[\nabla_{xx}f^*]^{-1}\,(\nabla_{xx}f-\nabla_{xx}f^*)\,[\nabla_{xx}f]^{-1}\,||$$

then we use again the strong convexity constant ℓ of assumption 2.4.4 to recover

$$||\,[\nabla_{xx}f^*]^{-1}\,(\nabla_{xx}f-\nabla_{xx}f^*)\,[\nabla_{xx}f]^{-1}\,||\leq\frac{1}{\ell^2}||(\nabla_{xx}f-\nabla_{xx}f^*)|| \tag{A.16}$$

now we apply the mean value theorem with $\tilde{x}$ as a point on the line between x_k and $x^*(t_k)$ to obtain

$$||(\nabla_{xx}f-\nabla_{xx}f^*)||\leq||\nabla_{xxx}f_{t_k}(\tilde{x})||\;||x_k-x^*(t_k)||\leq C_1||x_k-x^*(t_k)|| \tag{A.17}$$

Applying the same argument for the mixed second-order term implies

$$||(\nabla_{tx}f-\nabla_{tx}f^*)||\leq C_2||x_k-x^*(t_k)|| \tag{A.18}$$

Substituting A.17 and A.18 back into A.16 yields

$$||\,[\nabla_{xx}f]^{-1}\,\nabla_{xt}f-[\nabla_{xx}f^*]^{-1}\,\nabla_{xt}f^*||\leq\left(\frac{C_0C_1}{\ell^2}+\frac{C_2}{\ell}\right)||x_k-x^*(t_k)||$$

We have thus bounded all three terms of inequality A.15, substituting the three bounds in A.15 we remain with

$$||x_{k+1|k}-x^*(t_{k+1})||\leq\sigma||x_k-x^*(t_k)||+\frac{h^2}{2}\left[\frac{C_0^2C_1}{\ell^3}+\frac{2C_0C_2}{\ell^2}+\frac{2C_3}{\ell}\right] \tag{A.19}$$

where $\sigma:=1+h(C_0C_1/\ell^2+C_2/\ell)$ as defined in the statement of the theorem.
For the correction step 2.7 we may use the standard property of gradient descent for strongly convex functions with Lipschitz gradients. In particular, the Euclidean error norm of the gradient descent method converges as

$$||\hat{x}^{s+1}_{k+1}-x^*(t_{k+1})||\leq\rho||\hat{x}^{s}_{k+1}-x^*(t_{k+1})||$$

this is just proposition 2.2.1. Recalling that the sequence $\hat{x}^s_{k+1}$ is initialized by the predicted variable $x_{k+1|k}$ and the corrected variable x_{k+1} is equal to $\hat{x}^\tau_{k+1}$ we can write

$$||x_{k+1}-x^*(t_{k+1})||\leq\rho^\tau||x_{k+1|k}-x^*(t_{k+1})|| \tag{A.20}$$

we are now ready to consider the combined error bound achieved by the prediction correction scheme. By plugging the correction error of A.20 into the prediction error of A.19 we obtain

$$||x_{k+1}-x^*(t_{k+1})||\leq\rho^\tau\sigma||x_k-x^*(t_k)||+\rho^\tau\Gamma \tag{A.21}$$

where $\Gamma := h^2/2\,[C_0^2 C_1/\ell^3 + 2C_0 C_2/\ell^2 + 2C_3/\ell]$.
Notice that the relation A.21 between $||x_{k+1} - x^*(t_{k+1})||$ and $||x_k - x^*(t_k)||$ holds true even between $||x_k - x^*(t_k)||$ and $||x_{k-1} - x^*(t_{k-1})||$, i.e.,

$$||x_k - x^*(t_k)|| \leq \rho^\tau \sigma ||x_{k-1} - x^*(t_{k-1})|| + \rho^\tau \Gamma \tag{A.22}$$

So recursively applying the equation A.22 backwards in time to the initial time sample x_0 results in

$$||x_k - x^*(t_k)|| \leq (\rho^\tau \sigma)^k ||x_0 - x^*(t_0)|| + \rho^\tau \Gamma \sum_{i=0}^{k-1} (\rho^\tau \sigma)^i$$

if the time interval h is such that $\rho^\tau \sigma < 1$ this leads to

$$||x_k - x^*(t_k)|| \leq (\rho^\tau \sigma)^k ||x_0 - x^*(t_0)|| + \rho^\tau \Gamma \left[\frac{1 - (\rho^\tau \sigma)^k}{1 - \rho^\tau \sigma}\right]$$

this is the second statement of proposition 2.4.8.
To establish the other statement we can upper bound the term $||\left[\nabla_{xx} f\right]^{-1} \nabla_{xt} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{xt} f^*||$ in a worst case scenario utilizing the bound given in assumption 2.4.5 to obtain

$$||\left[\nabla_{xx} f\right]^{-1} \nabla_{xt} f - \left[\nabla_{xx} f^*\right]^{-1} \nabla_{xt} f^*|| \leq \frac{2C_0}{\ell}$$

substituting this bound into A.15 instead of the previously used bound yields

$$||x_{k+1|k} - x^*(t_{k+1})|| \leq ||x_k - x^*(t_k)|| + h\frac{2C_0}{\ell} + \Gamma$$

from the relation in A.20 this becomes

$$||x_{k+1} - x^*(t_{k+1})|| \leq \rho^\tau ||x_k - x^*(t_k)|| + \rho^\tau \left[h\frac{2C_0}{\ell} + \Gamma\right]$$

recursively iterating this equation backwards in time and using the standard formula of the geometric series we obtain the first statement of proposition 2.4.8

$$||x_{k+1} - x^*(t_{k+1})|| \leq \rho^{\tau(k+1)} ||x_k - x^*(t_k)|| + \rho^\tau \left[h\frac{2C_0}{\ell} + \Gamma\right] \left[\frac{1 - \rho^{\tau(k+1)}}{1 - \rho^\tau}\right]$$

Proof of proposition 2.4.9
Starting from distance between x_{t+1} and x^*_{t+1}.

$$\begin{aligned}
&\left|x_{t+1} - x^*_{t+1}\right| = |\theta x_t + \omega - \alpha \nabla f_{t+1}(\omega + \theta x_t) - \theta x^*_t - \omega| = \\
&|\theta(x_t - x^*_t) - \alpha(\nabla f_{t+1}(\omega + \theta x_t) - \nabla f_{t+1}(x^*_{t+1}))| = \\
&|\theta(x_t - x^*_t) - \alpha A_{t+1}(\omega + \theta x_t - x^*_{t+1})| \leq |\theta - \alpha\theta A_{t+1}|\,|x_t - x^*_t|
\end{aligned}$$

where we have used the notation A_{t+1} to denote the matrix (in this case a scalar) resulting from the $t+1$-th application of lemma 2.1.3. So we only need to guarantee that $|\theta - \alpha\theta A_{t+1}| < 1$. For this to be true we impose $(\theta-1)/\theta\ell < \alpha < (\theta+1)/\theta L$ using the fact that $\ell \leq A_t \leq Ł$ for all t by assumption. The bound $\theta < (\ell + L)/(L-\ell)$ guarantees that such an α exists.

Proof of proposition 2.4.10

Starting from distance between x_{t+1} and x^*_{t+1}.

$$\begin{aligned}&\left|x_{t+1} - x^*_{t+1}\right| = |\theta x_t + \omega - \alpha\nabla f_t(x_t) - \theta x^*_t - \omega| = \\ &|\theta(x_t - x^*_t) - \alpha(\nabla f_t(x_t) - \nabla f_t(x^*_t))| \leq |\theta - \alpha A_t|\,|x_t - x^*_t|\end{aligned}$$

where we have used the notation A_t to denote the matrix (in this case a scalar) resulting from the t-th application of lemma 2.1.3. So we only need to guarantee that $|\theta - \alpha A_t| < 1$. For this to be true we impose $(\theta-1)/\ell < \alpha < (\theta+1)/L$ as in the proof of proposition 2.4.9 having that $\ell \leq A_t \leq Ł$ for all t by assumption. The bound $\theta < (\ell+L)/(L-\ell)$ guarantees that such an α exists.

All the following proofs are original.

Proof of proposition 2.4.11

Starting from the distance between x_{t+1} and x^*_{t+1} and utilizing the mean value theorem we have

$$\begin{aligned}&|x_{t+1} - x^*_{t+1}| = |\phi(x_t) - \alpha\nabla f_t(x_t) - \phi(x^*_t)| = \\ &|\phi'(\bar{x})(x_t - x^*_t) - \alpha(\nabla f_t(x_t) - \nabla f_t(x^*_t))| \leq |\phi'(\bar{x}) - \alpha A_t|\,|x_t - x^*_t|\end{aligned}$$

where $\bar{x}$ is between x_t and x^*_t and A_t is the same as in proposition 2.4.10 and 2.4.9. So we only need to guarantee that $|\phi'(\bar{x}) - \alpha A_t| < 1$, from the bounds on ϕ' we need $(c+\epsilon-1)/\ell < \alpha < (c-\epsilon+1)/L$, the assumption $c - \epsilon < \phi'(\bar{x}) < c + \epsilon$ combined with $0 < \epsilon < 1$ and $0 < c < ((L+\ell)(-\epsilon+1))/(L-\ell)$ ensures that such an α exists.

Proof of proposition 2.4.12

Take

$$\begin{aligned}&\|x_{t+1} - x^*_{t+1}\| = \|\phi(x_t) - \alpha\nabla f_t(x_t) - \phi(x^*_t)\| = \\ &\|B(x_t - x^*_t) - \alpha(\nabla f_t(x_t) - \nabla f_t(x^*_t))\| \leq \|B - \alpha A_t\|\|x_t - x^*_t\|\end{aligned}$$

where we have called B the symmetric $d \times d$ matrix given by an application of lemma 2.1.3 to ϕ and A_t is as in propositions 2.4.9 and 2.4.10. For every sum of Hermitian matrices A, B it is true that the eigenvalues $\lambda_i(A), \lambda_i(B)$, that we assume ordered from smallest to largest, obey the inequality

$$\lambda_i(A+B) \leq \lambda_i(A) + \lambda_i(B)$$

and since $c - \epsilon \leq \lambda_i(B) \leq c + \epsilon$ and $\ell \leq \lambda_i(A) \leq L$ for all i choosing $(c+\epsilon-1)/\ell < \alpha < (c-\epsilon+1)/L$ as before, along with the conditions $0 < \epsilon < 1$ and $0 < c < ((L+\ell)(-\epsilon+1))/(L-\ell)$ that guarantee that such an α exists, will lead to $\|B - \alpha A_t\| < 1$ that concludes the proof.

Proof of proposition 2.4.12

Take

$$\left(x_{t+1} - x^*_{t+1}\right) = \sum_{i=1}^{p+1} \theta_i (x_{t+1-i} - x^*_{t+1-i}) - \sum_{i=1}^{p+1} \alpha_i \nabla f(x_{t+1-i}) =$$
$$\sum_{i=1}^{p+1} (\theta_i - \alpha_i A_i)(x_{t+1-i} - x^*_{t+1-i})$$

So we have obtained a recurrent sequence of the form

$$u_{t+1} = q_0 u_t + q_1 u_{t-1} + \cdots + q_p u_{t-p}$$

where $u_t = \left(x_{t+1} - x^*_{t+1}\right)$ and $q_i = (\theta_i - \alpha_i A_i)$. Choosing the α_i to make this recurrent sequence have all roots of its associated characteristic polynomial inside the unit circle will guarantee convergence. Notice that the α_i exist since there is a continuous function that links roots of a polynomial to its coefficients, and the coefficients θ_i are those of a recurrent sequence for which the characteristic equation has roots inside the unit circle. Appropriate bounds on the α can be found for specific values of p

Proof of proposition 2.4.13

As in proposition 2.2.1

$$||x_{k+1} - x^*_{k+1}|| = ||\omega + \beta x_k - \alpha \nabla f_k(x_k) - (\omega + \beta x^*_k + \epsilon_t)|| \leq$$
$$||I\beta - \alpha A_k||||x_k - x^*_k|| + ||\epsilon_t||$$

and we have that

$$\|I\beta - \alpha A_k\| < 1 \iff \frac{-1+\beta}{\ell} < \alpha < \frac{1+\beta}{\mathrm{L}}$$

as in proposition 2.4.10. So taking the expectation with respect to ϵ_t we can apply lemma 2.4.1 to reach the conclusion.

Proof of proposition 2.4.15

The sequence $\{f_t\}_{t\in\mathbb{N}}$ is just a sequence of random translations of the function f_0, so we immediately recover that for any t the function f_t is under assumption 2.0.1 and 2.0.2. Performing a quadratic Taylor expansion of $f_t(x_{t+1}+\epsilon_t)$ around x_t and using the fact that ∇f_t is Lipschitz $\forall t$ we obtain

$$\begin{aligned}
&f_{t+1}(x_{t+1}) = f_t(x_{t+1}+\epsilon_t) \leq \\
&f_t(x_t) + \nabla f_t(x_t)^T(x_{t+1}+\epsilon_t - x_t) + \frac{1}{2}\nabla^2 f_t(x_t)||x_{t+1}+\epsilon_t - x_t||_2^2 \\
&\leq f_t(x_t) + \nabla f_t(x_t)^T(x_{t+1}+\epsilon_t - x_t) + \frac{1}{2}\mathrm{L}||x_{t+1} - x_t||_2^2 + ||\epsilon_t||_2^2
\end{aligned}$$

taking expectations with respect to ϵ_t on both sides

$$\mathrm{E}_{\epsilon_t}[f_{t+1}(x_{t+1}) - f_t(x_t)] \leq \nabla f_t(x_t)^T(x_{t+1} - x_t) + \frac{1}{2}\mathrm{L}||x_{t+1} - x_t||_2^2 + \sigma_t^2$$

now we substitute in $x_{t+1} = x_t - \alpha\nabla f_{t+1}(x_t)$, the gradient descent update, to find

$$\mathrm{E}_{\epsilon_t}[f_{t+1}(x_{t+1}) - f_t(x_t)] \leq -(\alpha - \frac{1}{2}\mathrm{L}\alpha^2)||\nabla f_t(x_t)||_2^2 + M$$

where we have used that $\sigma_t^2 < M$. By lemma 2.1.5 we have that

$$2\ell(f_t(x_t) - f^*) \leq ||\nabla f_t(x_t)||_2^2 \tag{A.23}$$

where we are using the notation $f^* := f_t(x^*)$ we also notice that the minimum f^* of f_t does not depend on t since translations of a function don't change the value of the minimum.

Utilizing inequality A.23 and the assumption $\alpha < 1/\mathrm{L}$ that implies $\alpha(1/2 - L\alpha/2) \leq 0$ we have that

$$\begin{aligned}
&\mathrm{E}_{\epsilon_t}[f_{t+1}(x_{t+1}) - f_t(x_t)] \leq -(\alpha - \frac{1}{2}\mathrm{L}\alpha^2)||\nabla f_t(x_t)||_2^2 + M \leq \\
&-\frac{1}{2}\alpha||\nabla f_t(x_t)||_2^2 + M \leq -\ell\alpha(f_t(x_t) - f^*) + M
\end{aligned}$$

subtracting f^* from both sides and rearranging gives

$$\mathrm{E}_{\epsilon_t}[f_{t+1}(x_{t+1}) - f^*] \leq (1-\ell\alpha)(f_t(x_t) - f^*) + M$$

and taking expectations with respect to the joint distribution of all $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_1$ yields

$$\mathrm{E}[f_{t+1}(x_{t+1}) - f^*] \leq (1-\ell\alpha)\mathrm{E}[f_t(x_t) - f^*] + M$$

where the notation on the expected value is left free to avoid cluttering. Subtracting the constant $M/\ell\alpha$ from both sides one obtains

$$\mathrm{E}[f_{t+1}(x_{t+1}) - f^*] - \frac{M}{\ell\alpha} \leq (1 - \ell\alpha)\left(\mathrm{E}[f_t(x_t) - f^*] - \frac{M}{\ell\alpha}\right)$$

this is a contraction inequality since $\alpha\ell \leq \ell/\mathrm{L} \leq 1$. So we can recursively apply it and the result will follow.

Appendix B

All proofs in this appendix aside from the first one are original.

Proof of proposition 4.1.5

By repeated application of the mean value theorem to $\tilde{p}(y|\tilde{\lambda}_{t+1}, \boldsymbol{\theta})$ and $\tilde{s}_t(\tilde{\lambda}_{t+1}, y_t, \boldsymbol{\theta})$ and using the form of the Newton-score update 4.2, we manage to obtain locally realized Kullback-Leibler optimality by starting from the definition

$$\begin{aligned}
&\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \log \frac{\tilde{p}(y|\tilde{\lambda}_t, \boldsymbol{\theta})}{(y|\tilde{\lambda}_{t+1}, \boldsymbol{\theta})} dy = \\
&-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \log \frac{\partial \tilde{p}(y|\tilde{\lambda}^*_{t+1}, \boldsymbol{\theta})}{\partial \tilde{\lambda}} (\tilde{\lambda}_{t+1} - \tilde{\lambda}_t) dy = \\
&-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \tilde{s}_t(\tilde{\lambda}^*_{t+1}, y_t, \boldsymbol{\theta}) \alpha S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) dy = \\
&-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \alpha S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \left(\tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta})\right)^2 dy \qquad \text{(B.1)} \\
&-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \alpha S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) \frac{\partial \tilde{s}_t(\tilde{\lambda}^{**}_t, y^{**}_t, \boldsymbol{\theta})}{\partial y} (y_t - y) dy \\
&-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \alpha S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta}) \frac{\partial \tilde{s}_t(\tilde{\lambda}^{**}_t, y^{**}_t, \boldsymbol{\theta})}{\partial \tilde{\lambda}} (\tilde{\lambda}^*_{t+1} - \tilde{\lambda}_t) dy := \\
&-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t) \alpha S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \left(\tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta})\right)^2 dy + a(\delta_\lambda, \delta_y) + b(\delta_\lambda, \delta_y) < 0 \qquad \text{(B.2)}
\end{aligned}$$

where $\tilde{\lambda}^*_{t+1}$ is a point between $\tilde{\lambda}_{t+1}$ and $\tilde{\lambda}_t$, $\tilde{\lambda}^{**}_{t+1}$ is a point between $\tilde{\lambda}^*_{t+1}$ and $\tilde{\lambda}_t$, y^{**}_t is a point between y_t and y, and $a(\delta_\lambda, \delta_y)$, $b(\delta_\lambda, \delta_y)$ in B.2 are equal to the second and third remainders terms of B.1, respectively. From assumption 4.1.3 and 4.1.4 we obtain that $\alpha S_t(\tilde{\lambda}_t, \boldsymbol{\theta}) \left(\tilde{s}_t(\tilde{\lambda}_t, y_t, \boldsymbol{\theta})\right)^2 > 0$ almost surely, so for every $\tilde{\lambda}_t$ and p_t there exists $\gamma < 0$

such that

$$-\int_{Y_{\delta_y}(y_t)} p(y|\lambda_t)\alpha S_t(\tilde{\lambda}_t,\boldsymbol{\theta})\left(\tilde{s}_t(\tilde{\lambda}_t,y_t,\boldsymbol{\theta})\right)^2 dy \leq \gamma < 0$$

The desired result now follows upon noting that the second and third terms in B.1 can be made arbitrarily small compared to the first term, due to the differentiability of the score and the compactness of $Y_{\delta_y}(y_t)$. Notice also that $a(\delta_\lambda,\delta_y), b(\delta_\lambda,\delta_y)$ go to zero quicker than the first term in equation B.1 since both their domain of integration and the function to integrate go to zero.

Proof of proposition 4.2.3

From the square of the norm we have

$$\begin{aligned}&\|E_{Y_t|Y^{t-1}}[\tilde{\lambda}_{t+1}] - \lambda_t^*\|^2 = \|\tilde{\lambda}_t - \alpha S_t\nabla f_t(\tilde{\lambda}_t) - \lambda_t^*\|^2 = \\ &\|\tilde{\lambda}_t - \lambda_t^*\|^2 - 2\alpha S_t\nabla f_t(\tilde{\lambda}_t)(\tilde{\lambda}_t - \lambda_t^*) + \alpha^2 S_t^2\|\nabla f_t(\tilde{\lambda}_t)\|^2.\end{aligned}$$

where we have used assumption 4.2.2 in the first equality.
Since assumptions 2.0.1, 2.0.2 are satisfied we can use lemma 2.1.7 (the co-coercivity of the convex objective function $f_t(\tilde{\lambda}_t)$) to obtain

$$\begin{aligned}&\|\tilde{\lambda}_t - \lambda_t^*\|^2 - 2\alpha S_t\nabla f_t(\tilde{\lambda}_t)(\tilde{\lambda}_t - \lambda_t^*) + \alpha^2 S_t^2\|\nabla f_t(\tilde{\lambda}_t)\|^2 \leq \\ &\|\tilde{\lambda}_t - \lambda_t^*\|^2 - \alpha S_t\left(\frac{2}{L} - \alpha S_t\right)\|\nabla f_t(\tilde{\lambda}_t)\|^2.\end{aligned}$$

Thus, from the assumption that $0 < S_t\alpha < 2/L$, the result follows.

Proof of lemma 4.2.6

From assumption 4.2.2

$$\nabla E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t,\boldsymbol{\theta})\nabla\log\tilde{p}(Y_t|\tilde{\lambda}_t,\boldsymbol{\theta})] = \int_{\mathbb{R}} p_t(y_t|\lambda_t)\nabla\left(S_t(\tilde{\lambda}_t,\boldsymbol{\theta})\nabla\log\tilde{p}(y_t|\tilde{\lambda}_t,\boldsymbol{\theta})\right) dy_t$$

Notice that the conditioning allows us to treat $\tilde{\lambda}_t$ like a constant since it only depends on the previous random variables Y^{t-1}. Then, since assumption 4.2.4 holds uniformly over all y_t, we have that

$$-\int_{\mathbb{R}} p_t(y_t|\lambda_t)\nabla\left(S_t(\tilde{\lambda}_t,\boldsymbol{\theta})\nabla\log\tilde{p}(y_t|\tilde{\lambda}_t,\boldsymbol{\theta})\right) dy_t \leq \int_{\mathbb{R}} p_t(y_t|\lambda_t)IL\; dy_t = LI.$$

Utilizing assumption 4.2.5, the same argument holds true for the inequality regarding ℓ. The fact that the critical point λ_t^* remains the only minimum individuated by the roots of the equality

$$E_{Y_t|Y^{t-1}}[S_t(\lambda_t^*, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] = S_t(\lambda_t^*, \boldsymbol{\theta})E_{Y_t|Y^{t-1}}[\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] = 0$$

follows from the previous two inequalities, that identify it as the unique maximum of a strongly convex function.

Proof of proposition 4.2.7

As in lemma 2.2.2 we know that $E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})]$ is a Lipschitz continuous gradient of a strongly convex function that shares the same minimum λ_t^* of $f_t(\tilde{\lambda})$. Moreover from the fundamental theorem of calculus we can write

$$\begin{aligned}
&E_{Y_t|Y^{t-1}}[S_t(\lambda_t^*, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] - E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] = \\
&\int_0^1 \frac{\partial E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t + \tau(\lambda_t^* - \tilde{\lambda}_t), \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t + \tau(\lambda_t^* - \tilde{\lambda}_t, \boldsymbol{\theta})]}{\partial \tau} d\tau = \\
&\int_0^1 \nabla E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t + \tau(\lambda_t^* - \tilde{\lambda}_t), \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t + \tau(\lambda_t^* - \tilde{\lambda}_t, \boldsymbol{\theta})](\lambda_t^* - \tilde{\lambda}_t) d\tau
\end{aligned}$$

from which we obtain that

$$E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t^*, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] = E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] + A_t(\lambda_t^* - \tilde{\lambda}_t)$$

where

$$A_t := \int_0^1 \nabla E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t + \tau(\lambda_t^* - \tilde{\lambda}_t), \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t + \tau(\lambda_t^* - \tilde{\lambda}_t, \boldsymbol{\theta})] d\tau$$

by assumption we thus have $I_{n \times n}\ell \leq A_t \leq I_{n \times n}L$. Notice how we needed the conditioning because knowing Y^{t-1} the model time varying parameter $\tilde{\lambda}_t$ behaves like a constant inside the conditional expectation. Following the proof idea of proposition 2.2.1 we have

$$\begin{aligned}
&||E_{Y_t|Y^{t-1}}[\tilde{\lambda}_{t+1}] - \lambda_t^*|| = ||\tilde{\lambda}_t + \alpha E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] - \lambda_t^*|| = \\
&||\tilde{\lambda}_t + \alpha E_{Y_t|Y^{t-1}}[S_t(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] - \alpha E_{Y_t|Y^{t-1}}[S_t(\lambda_t^*, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] - \lambda_t^*|| \leq \\
&||(\tilde{\lambda}_t - \lambda_t^*)(I - \alpha A_t)|| \leq ||(I - \alpha A_t)||\ ||(\tilde{\lambda}_t - \lambda_t^*)||
\end{aligned}$$

where we have used lemma 4.2.6 to add $E_{Y_t|Y^{t-1}}[S_t(\lambda_t^*, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] = 0$. For every symmetric matrix A_t we have $||I - A_t|| \leq \max\{|1 - \alpha\xi_1|, |1 - \alpha\xi_n|\}$ where ξ_1, ξ_n are respectively the smallest and the largest eigenvalues of A_t. Hence $||\tilde{\lambda}_{t+1} - \lambda_t^*|| \leq$

$q||\tilde{\lambda}_t - \lambda_t^*||$ where $q = \max\{|1-\alpha\ell|, |1-\alpha L|\}$. Since $0 < \alpha < 2/L,\ 0 < \ell \leq L$ then $|1-\alpha\ell| < 1, |1-\alpha L| < 1$ that means $q < 1$.

Proof of proposition 4.2.13

We have

$$||E_{Y_t|Y^{t-1}}[\tilde{\lambda}_{t+1}] - \lambda_t^*||^2 = ||E[\tilde{\lambda}_t] - \alpha E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] - \lambda_t^*||^2$$

and we know that $E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})]$ behaves like the gradient of a strictly convex function with Lipschitz continuous derivative by the same argument of lemma 4.2.6 (here we use assumption 4.2.2 and 4.2.4). Thus all conditions to apply lemma 2.1.7 are met so we use the co-coercivity of convex functions to deduce that

$$||\tilde{\lambda}_t - \alpha E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] - \lambda_t^*||^2 = ||\tilde{\lambda}_t - \lambda_t^*||^2 -$$
$$2\alpha(\tilde{\lambda}_t - \lambda_t^*)^T E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] + \alpha^2 ||E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})]||^2 <$$
$$||\tilde{\lambda}_t - \lambda_t^*||^2 - \alpha\left(\frac{2}{L} - \alpha\right) ||E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})]||^2$$

so from the fact that $\alpha < \frac{2}{L}$ we obtain

$$||E_{Y_t|Y^{t-1}}[\tilde{\lambda}_{t+1}] - \lambda_t^*|| \leq ||\tilde{\lambda}_t - \lambda_t^*||.$$

Proof of proposition 4.2.16

As in proposition 2.2.1

$$\|E_{Y_t|Y^{t-1}}[\tilde{\lambda}_{t+1}] - E_{\epsilon_t}[\lambda_{t+1}^*]\| = \|\omega + \beta\tilde{\lambda}_t - \alpha E_{Y_t|Y^{t-1}}[S(\tilde{\lambda}_t, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\tilde{\lambda}_t, \boldsymbol{\theta})] - \omega - \beta\lambda_t^*\| =$$
$$\|(I\beta - \alpha A_t)(\tilde{\lambda}_t - \lambda_t^*)\| \leq \|I\beta - \alpha A_t\| \|\tilde{\lambda}_t - \lambda_t^*\|$$

where we have used 4.2.6 to add $E_{Y_t|Y^{t-1}}[S_t(\lambda_t^*, \boldsymbol{\theta})\nabla \log \tilde{p}(Y_t|\lambda_t^*, \boldsymbol{\theta})] = 0$ and the same argument as in proposition 4.2.7 is followed (the assumption are used here).
Then for every symmetric matrix A_t we have $||I\beta - A_t|| \leq \max\{|\beta - \alpha\xi_1|, |\beta - \alpha\xi_n|\}$ where ξ_1, ξ_n are respectively the smallest and the largest eigenvalues of A_t. Hence $||\tilde{\lambda}_{t+1} - \lambda_t^*|| \leq q||\tilde{\lambda}_t - \lambda_t^*||$ where $q = \max\{|\beta - \alpha\ell|, |\beta - \alpha L|\}$ and Since $0 < \alpha < (\beta+1)/L$ then $|\beta - \alpha\ell| < 1, |\beta - \alpha L| < 1$ that in turn means $q < 1$.

Proof of corollary 4.2.17

We only need to show that assumptions 4.2.4 , 4.2.5 , 4.2.2 hold and then we can apply proposition 4.2.16 to reach the conclusion. Assumption 4.2.2 holds since

$$y_t \log \tilde{p}_t(y_t|\tilde{\sigma}^2, \mu) = -\frac{y_t}{2\tilde{\sigma}^2} + \frac{y_t(y_t-\mu)^2}{2\tilde{\sigma}^4}$$

is continuous in both variables y_t and $\tilde{\sigma}^2$ so an application of the Leibniz integral rule allows the interchange of integral and derivative.
While assumptions 4.2.4 and 4.2.5 hold for the same reasons as in example 4.2.9 .

Proof of proposition 4.2.19

The exact same argument of proposition 4.2.16 applies noticing that q will still be smaller than one.

Proof of corollary 4.2.19

We only need to show that assumptions 4.2.4, 4.2.2 and equation 4.5 hold and then we can apply proposition 4.2.18 to reach the conclusion. Assumption 4.2.2 holds since

$$y_t \log \tilde{p}_t(y_t|\tilde{\sigma}, \mu) = y_t \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\exp(\tilde{\sigma})}\right) - \frac{\nu+1}{2}\log\left(1 + \frac{1}{\nu}\left(\frac{y_t}{\exp(\tilde{\sigma})}\right)^2\right)$$

is continuous in both variables y_t and $\tilde{\sigma}$ so an application of the Leibniz integral rule allows the interchange of integral and derivative.
While assumptions 4.2.4 and 4.2.5 hold for the same reasons as in example 4.2.12.

Bibliography

H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998. ISBN 978-1-4612-1694-0. doi: 10.1007/978-1-4612-1694-0_15. URL https://doi.org/10.1007/978-1-4612-1694-0_15.

R. Alzghool. Parameters estimation for garch (p,q) model: Ql and aql approaches. *Electronic Journal of Applied Statistical Analysis*, 10:180–193, 01 2017. doi: 10.1285/i20705948v10n1p180.

Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4):185 – 196, 1993. ISSN 0925-2312. doi: https://doi.org/10.1016/0925-2312(93)90006-O. URL http://www.sciencedirect.com/science/article/pii/092523129390006O.

A. Ayala and S. Blazsek. Equity market neutral hedge funds and the stock market: an application of score-driven copula models. *Applied Economics*, 50(37):4005–4023, 2018. doi: 10.1080/00036846.2018.1440062. URL https://doi.org/10.1080/00036846.2018.1440062.

A. Babii, X. Chen, and E. Ghysels. Commercial and residential mortgage defaults: Spatial dependence with frailty. *Journal of Econometrics*, 212(1):47 – 77, 2019. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2019.04.020. URL http://www.sciencedirect.com/science/article/pii/S0304407619300752. Big Data in Dynamic Predictive Econometric Modeling.

R. A. Bates, R. J. Buck, E. Riccomagno, and H. P. Wynn. Experimental design and observation for large systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):77–94, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346166.

D. A. Belsley. *On the Determination of Systematic Parameter Variation in the Linear*

Regression Model, pages 487–494. NBER, October 1973. URL http://www.nber.org/chapters/c9939.

D. A. Belsley and E. Kuti. Time-Varying Parameter Structures: An Overview. In *Annals of Economic and Social Measurement, Volume 2, number 4*, NBER Chapters, pages 375–379. National Bureau of Economic Research, Inc, June 1973. URL https://ideas.repec.org/h/nbr/nberch/9932.html.

A. K. Bera and M. L. Higgins. Arch models: Properties, estimation and testing. *Journal of Economic Surveys*, 7(4):305–366. doi: 10.1111/j.1467-6419.1993.tb00170.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.1993.tb00170.x.

I. Berkes, L. Horváth, and P. Kokoszka. Garch processes: structure and estimation. *Bernoulli*, 9(2):201–227, 04 2003. doi: 10.3150/bj/1068128975. URL https://doi.org/10.3150/bj/1068128975.

A. Bernardi and M. Bernardi. *Two-Sided Skew and Shape Dynamic Conditional Score Models: MAF 2018*, pages 121–124. 07 2018. ISBN 978-3-319-89823-0. doi: 10.1007/978-3-319-89824-7_22.

D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*, volume 27. 01 1996. doi: 10.1007/978-0-387-74759-0_440.

K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, and B. Armstrong. Time series regression studies in environmental epidemiology. *International journal of epidemiology*, 42, 06 2013. doi: 10.1093/ije/dyt092.

Bittanti and F. Cuzzola. A mixed gh2/h approach for stabilization and accurate trajectory tracking of unicycle-like vehicles. *International Journal of Control*, 74:880–888, 06 2001.

F. Black. Studies of stock price volatility changes. 1976.

F. Blasques, S. Koopman, and A. Lucas. Maximum likelihood estimation for correctly specified generalized autoregressive score models: Feedback effects, contraction conditions and asymptotic properties. WorkingPaper 14-074/III, Tinbergen Institute, 2014a.

F. Blasques, S. J. Koopman, and A. Lucas. Maximum Likelihood Estimation for Score-Driven Models. Tinbergen Institute Discussion Papers 14-029/III, Tinbergen Institute, Mar. 2014b. URL https://ideas.repec.org/p/tin/wpaper/20140029.html.

F. Blasques, S. J. Koopman, and A. Lucas. Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 102(2): 325–343, 03 2015. ISSN 0006-3444. doi: 10.1093/biomet/asu076. URL https://doi.org/10.1093/biomet/asu076.

F. Blasques, P. Gorgi, S. J. Koopman, and O. Wintenberger. Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. Tinbergen Institute Discussion Paper 16-082/III, Amsterdam and Rotterdam, 2016a. URL http://hdl.handle.net/10419/149486.

F. Blasques, S. J. Koopman, A. Lucas, and J. Schaumburg. Spillover dynamics for systemic risk measurement using spatial financial time series models. *Journal of Econometrics*, 195(2):211 – 223, 2016b. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2016.09.001. URL http://www.sciencedirect.com/science/article/pii/S0304407616301609.

F. Blasques, A. Lucas, and A. Vlodrop. Finite sample optimality of score-driven volatility models. *SSRN Electronic Journal*, 01 2017. doi: 10.2139/ssrn.3076829.

F. Blasques, S. J. Koopman, and A. Lucas. Amendments and Corrections: 'Information-theoretic optimality of observation-driven time series models for continuous responses'. *Biometrika*, 105(3):753–753, 07 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy039. URL https://doi.org/10.1093/biomet/asy039.

F. Blasques, S. J. Koopman, and A. Lucas. Nonlinear autoregressive models with optimality properties. *Econometric Reviews*, 0(0):1–20, 2019. doi: 10.1080/07474938.2019.1701807. URL https://doi.org/10.1080/07474938.2019.1701807.

S. Blazsek and M. Villatoro. Is beta-t-egarch(1,1) superior to garch(1,1)? *Applied Economics*, 47(17):1764–1774, 2015. doi: 10.1080/00036846.2014.1000536. URL https://doi.org/10.1080/00036846.2014.1000536.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, apr 1986. doi: 10.1016/0304-4076(86)90063-1. URL https://doi.org/10.1016/0304-4076(86)90063-1.

T. Bollerslev. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69(3):542–547, 1987. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/1925546.

L. Bottou. *On-line Learning and Stochastic Approximations*, page 9–42. Publications of the Newton Institute. Cambridge University Press, 1999. doi: 10.1017/CBO9780511569920.003.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60:223–311, 2016.

L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

G. Box, G. Jenkins, G. Reinsel, and G. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN 9781118674925. URL https://books.google.nl/books?id=rNt5CgAAQBAJ.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

B. O. Bradley and M. S. Taqqu. Financial risk and heavy tails. In S. T. Rachev, editor, *Handbook of Heavy Tailed Distributions in Finance*, volume 1 of *Handbooks in Finance*, pages 35 – 103. North-Holland, Amsterdam, 2003. doi: https://doi.org/10.1016/B978-044450896-6.50004-2. URL http://www.sciencedirect.com/science/article/pii/B9780444508966500042.

L. Catania, S. Grassi, and F. Ravazzolo. *Predicting the Volatility of Cryptocurrency Time-Series: MAF 2018*, pages 203–207. 07 2018. ISBN 978-3-319-89823-0. doi: 10.1007/978-3-319-89824-7_37.

A. Cauchy. Méthode générale pour la résolution des systèmes d' equations simultanées. *C. R. Acad. Sci. Paris*, pages 536–538, 1847.

K. S. Chan and H. Tong. On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, 7(3):179–190, 1986. doi: 10.1111/j.1467-9892.1986.tb00501.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1986.tb00501.x.

T. F. Cooley and E. C. Prescott. Estimation in the presence of stochastic parameter variation. *Econometrica*, 44(1):167–184, 1976. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1911389.

F. Corsi. A simple long memory model of realized volatility. *Journal of Financial Econometrics*, 7:174–196, 02 2009. doi: 10.1093/jjfinec/nbp001.

D. R. Cox, G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen. Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 8(2):93–115, 1981. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615819.

D. Creal, S. J. Koopman, and A. Lucas. A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, 29(4):552–563, 2011. doi: 10.1198/jbes.2011.10070. URL https://doi.org/10.1198/jbes.2011.10070.

D. Creal, S. J. Koopman, and A. Lucas. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795, 2013. doi: 10.1002/jae.1279. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1279.

J. Davidson. A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes. *Econometric Theory*, 8(3):313–329, 1992. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/3532351.

S. Degiannakis and E. Xekalaki. Autoregressive conditional heteroscedasticity (arch) models: A review. *Quality Technology & Quantitative Management*, 1(2):271–324, 2004. doi: 10.1080/16843703.2004.11673078. URL https://doi.org/10.1080/16843703.2004.11673078.

E. M. Dogo, O. J. Afolabi, N. I. Nwulu, B. Twala, and C. O. Aigbavboa. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 92–99, Dec 2018. doi: 10.1109/CTEMS.2018.8769211.

R. Douc, P. Doukhan, and E. Moulines. Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications*, 123(7):2620 – 2647, 2013. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2013.04.010. URL http://www.sciencedirect.com/science/article/pii/S0304414913001051. A Special Issue on the Occasion of the 2013 International Year of Statistics.

J. Du. The frontier of SGD and its variants in machine learning. *Journal of Physics: Conference Series*, 1229:012046, may 2019. doi: 10.1088/1742-6596/1229/1/012046. URL https://doi.org/10.1088%2F1742-6596%2F1229%2F1%2F012046.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987, jul 1982. doi: 10.2307/1912773. URL https://doi.org/10.2307/1912773.

R. F. Engle and V. K. Ng. Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5):1749–1778, 1993. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/2329066.

R. F. Engle and J. R. Russell. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1162, 1998. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2999632.

R. Ferland, A. Latour, and D. Oraichi. Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942, 2006. doi: 10.1111/j.1467-9892.2006.00496.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.2006.00496.x.

J. Fleming and C. Kirby. A Closer Look at the Relation between GARCH and Stochastic Autoregressive Volatility. *Journal of Financial Econometrics*, 1(3):365–419, 09 2003. ISSN 1479-8409. doi: 10.1093/jjfinec/nbg016. URL https://doi.org/10.1093/jjfinec/nbg016.

R. V. Fonseca and F. Cribari-Neto. Bimodal birnbaum–saunders generalized autoregressive score model. *Journal of Applied Statistics*, 45(14):2585–2606, 2018. doi: 10.1080/02664763.2018.1428734. URL https://doi.org/10.1080/02664763.2018.1428734.

C. Francq and J.-M. Zakoian. Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, 10(4):605–637, 08 2004. doi: 10.3150/bj/1093265632. URL https://doi.org/10.3150/bj/1093265632.

C. Francq and J.-M. Zakoian. *GARCH models*. Wiley, New York, 2010.

A. Gallant. *Nonlinear statistical models*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987. ISBN 9780471802600. URL https://books.google.nl/books?id=6TPvAAAAMAAJ.

A. Gallant and H. White. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. B. Blackwell, 1988. ISBN 9780631157656. URL https://books.google.nl/books?id=VVOqQgAACAAJ.

E. G. Gladyshev. On stochastic approximation. *Theory of Probability & Its Applications*, 10(2):275–278, 1965. doi: 10.1137/1110031. URL https://doi.org/10.1137/1110031.

L. Glosten, R. Jagannanthan, and D. Runkle. On the relation between expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5): 1779–1802, 1993.

P. Gorgi, S. Koopman, and R. Lit. The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 04 2019. doi: 10.1111/rssa.12464.

R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik. Sgd: General analysis and improved rates, 2019.

P. R. Hansen and A. Lunde. A forecast comparison of volatility models: does anything beat a garch(1,1)? *Journal of Applied Econometrics*, 20(7):873–889, 2005a. doi: 10.1002/jae.800. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.800.

P. R. Hansen and A. Lunde. A forecast comparison of volatility models: does anything beat a garch(1,1)? *Journal of Applied Econometrics*, 20(7):873–889, 2005b. doi: 10.1002/jae.800. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.800.

A. Harvey and T. Chakravarty. Beta-t-(e)garch. Cambridge working papers in economics, Faculty of Economics, University of Cambridge, 2008. URL https://EconPapers.repec.org/RePEc:cam:camdae:0840.

A. Harvey and A. Luati. Filtering with heavy tails. *Journal of the American Statistical Association*, 109, 07 2014a. doi: 10.1080/01621459.2014.887011.

A. Harvey and A. Luati. Filtering with heavy tails. *Journal of the American Statistical Association*, 109(507):1112–1122, 2014b. doi: 10.1080/01621459.2014.887011. URL https://doi.org/10.1080/01621459.2014.887011.

A. C. Harvey. *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Econometric Society Monographs. Cambridge University Press, 2013. doi: 10.1017/CBO9781139540933.

I. A. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability & Its Applications*, 7(4):349–382, 1962. doi: 10.1137/1107036. URL https://doi.org/10.1137/1107036.

S. T. Jensen and A. Rahbek. Asymptotic inference for nonstationary garch. *Econometric Theory*, 20(6):1203–1226, 2004. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/3533451.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 315–323, Red Hook, NY, USA, 2013a. Curran Associates Inc.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 315–323, Red Hook, NY, USA, 2013b. Curran Associates Inc.

S. M. Kakade and S. Shalev-Shwartz. On the duality of strong convexity and strong smoothness : Learning applications and matrix regularization. 2009.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-undefinedojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, page 795–811, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783319461274. doi: 10.1007/978-3-319-46128-1_50. URL https://doi.org/10.1007/978-3-319-46128-1_50.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

R. Kleinberg, Y. Li, and Y. Yuan. An alternative view: When does SGD escape local minima? *CoRR*, abs/1802.06175, 2018. URL http://arxiv.org/abs/1802.06175.

A. Knutson and T. Tao. Honeycombs and sums of hermitian matrices. 2001.

P. Koch, T. Simpson, J. Allen, and F. Mistree. Statistical approximations for multidisciplinary design optimization: The problem of size. *Journal of Aircraft*, 36:275–286, 01 1999. doi: 10.2514/2.2435.

S. J. Koopman, A. Lucas, and M. Scharth. Predicting time-varying parameters with parameter-driven and observation-driven models. *The Review of Economics and Statistics*, 98(1):97–110, 2016. doi: 10.1162/REST_a_00533. URL https://doi.org/10.1162/REST_a_00533.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

S.-W. Lee and B. Hansen. Asymptotic theory for the garch(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10(1):29–52, 1994a. URL https://EconPapers.repec.org/RePEc:cup:etheor:v:10:y:1994:i:01:p:29-52_00.

S.-W. Lee and B. E. Hansen. Asymptotic theory for the garch(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10(1):29–52, 1994b. doi: 10.1017/S0266466600008215.

Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, PP, 12 2019. doi: 10.1109/TNNLS.2019.2952219.

R. Lumsdaine. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in igarch(1,1) and covariance stationary garch(1,1) models. *Econometrica*, 64:575–96, 02 1996a. doi: 10.2307/2171862.

R. L. Lumsdaine. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in igarch(1,1) and covariance stationary garch(1,1) models. *Econometrica*, 64(3):575–96, 1996b. URL https://EconPapers.repec.org/RePEc:ecm:emetrp:v:64:y:1996:i:3:p:575-96.

M. Métivier. *Semimartingales.* De Gruyter, Berlin, Boston, 2011. ISBN 978-3-11-084556-3.

A. Monfort. A reappraisal of misspecified econometric models. *Econometric Theory*, 12 (4):597–619, 1996. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/3532787.

D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2938260.

Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

L. Nguyen, P. Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, and M. Takáč. Sgd and hogwild! convergence without the bounded gradients assumption. 07 2018.

R. S. Pindyck and D. L. Rubinfeld. *Econometric models and economic forecasts*. Boston, Mass. : Irwin/McGraw-Hill, 4th ed edition, 1998. ISBN 0071158367 (pbk.). Includes indexes.

B. T. Polyak. *Introduction to Optimization (Translations Series in Mathematics and Engineering)*. Optimization Software, 1987. ISBN 0911575146. URL https://www.amazon.com/Introduction-Optimization-Translations-Mathematics-Engineering/dp/0911575146?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0911575146.

A. Y. Popkov. Gradient methods for nonstationary unconstrained optimization problems. *Automation and Remote Control*, 66(6):883–891, Jun 2005. ISSN 1608-3032. doi: 10.1007/s10513-005-0132-z. URL https://doi.org/10.1007/s10513-005-0132-z.

N. Qian. On the momentum term in gradient descent learning algorithms. *Neural Netw.*, 12(1):145–151, Jan. 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00116-6. URL https://doi.org/10.1016/S0893-6080(98)00116-6.

R. R. Rao. Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(2):659–680, 1962. ISSN 00034851. URL http://www.jstor.org/stable/2237541.

S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. J. Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in neural information processing systems*, pages 2647–2655, 2015.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. ISSN 00034851. URL http://www.jstor.org/stable/2236626.

J. R. Russell. Econometric modeling of multivariate irregularly-spaced high-frequency data. 2000.

J. R. Russell and R. F. Engle. A discrete-state continuous-time model of financial transactions prices and times. *Journal of Business & Economic Statistics*, 23(2): 166–180, 2005. doi: 10.1198/073500104000000541. URL https://doi.org/10.1198/073500104000000541.

A. H. Sarris. A Bayesian Approach To Estimation Of Time-Varying Regression Coefficients. In *Annals of Economic and Social Measurement, Volume 2, number 4*, NBER

Chapters, pages 501–523. National Bureau of Economic Research, Inc, June 1973. URL https://ideas.repec.org/h/nbr/nberch/9941.html.

N. Shephard. Statistical aspects of arch and stochastic volatility, in cox, dr, dv hinkley and oe barndorff0nielsen (eds), time series models in econometrics, finance and other fields, 1996.

A. Simonetto and E. Dall'Anese. Prediction-correction algorithms for time-varying constrained optimization. *IEEE Transactions on Signal Processing*, 65(20):5481–5494, Oct 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2728498.

A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro. A class of prediction-correction methods for time-varying convex optimization. *IEEE Transactions on Signal Processing*, 64(17):4576–4591, Sep. 2016. ISSN 1053-587X. doi: 10.1109/TSP.2016.2568161.

E. Slutsky. The summation of random causes as the source of cyclic processes. *Moscow: Conjuncture Institute*, 1927, 1927.

A. Sorokin. Non-invertibility in some heteroscedastic models. *SSRN Electronic Journal*, 04 2011. doi: 10.2139/ssrn.1908661.

D. Strauman and T. Mikosch. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Annals of Statistics*, 34, 03 2006. doi: 10.1214/009053606000000803.

H. Tong and K. S. Lim. *Threshold Autoregression, Limit Cycles and Cyclical Data*, pages 9–56. 2009. doi: 10.1142/9789812836281_0002. URL https://www.worldscientific.com/doi/abs/10.1142/9789812836281_0002.

G. T. Walker. On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 131(818):518–532, 1931.

H. White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25, January 1982. URL https://ideas.repec.org/a/ecm/emetrp/v50y1982i1p1-25.html.

O. Wintenberger. Continuous invertibility and stable qml estimation of the egarch(1,1) model. *Scandinavian Journal of Statistics*, 40(4):846–867, 2013. doi: 10.1111/sjos.12038. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12038.

G. Yule. On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London Series A*, 226:267–298, Jan. 1927.

S. L. Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988. ISSN 00063444. URL http://www.jstor.org/stable/2336303.

E. Zivot. *Practical Issues in the Analysis of Univariate GARCH Models*, pages 113–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71297-8. doi: 10.1007/978-3-540-71297-8_5. URL https://doi.org/10.1007/978-3-540-71297-8_5.