Statistical Papers
https://doi.org/10.1007/s00362-023-01497-y

REGULAR ARTICLE

Maximum Likelihood With a Time Varying Parameter

Alberto Lanconelli[1] · **Christopher S. A. Lauria**[1]

Received: 27 February 2023 / Revised: 29 August 2023

Abstract
We consider the problem of tracking an unknown time varying parameter that characterizes the probabilistic evolution of a sequence of independent observations. To this aim, we propose a stochastic gradient descent-based recursive scheme in which the log-likelihood of the observations acts as time varying gain function. We prove convergence in mean-square error in a suitable neighbourhood of the unknown time varying parameter and illustrate the details of our findings in the case where data are generated from distributions belonging to the exponential family.

Keywords Stochastic gradient descent · Maxim likelihood · Exponential family

Mathematics Subject Classification 65K05 · 62F12

1 Introduction

When estimating unknown parameters in a dynamic model the optimum solution to the parameter estimation problem may not remain constant. Specifically, the optimal values of the model parameters may change through time because of the evolution of the underlying process: finding them is, in general, not straightforward. A survey of basic techniques for tracking the time-varying dynamics of a system is provided in Ljung and Gunnarsson (1990) where recursive algorithms in non-stationary stochastic optimization are analysed under different assumptions about the true system's variations, see also Simonetto et al. (2020) for a review in a purely deterministic setting. In Delyon and Juditsky (1995) the problem of tracking the random drifting parameters of a linear regression system is tackled, and Zhu and Spall (2016) builds a computable tracking error bound for how a stochastic approximation with constant gain keeps up

✉ Alberto Lanconelli
alberto.lanconelli2@unibo.it

Christopher S. A. Lauria
christopher.lauria2@unibo.it

[1] Dipartimento di Scienze Statistiche Paolo Fortunati, Università di Bologna, Bologna, Italy

Published online: 29 September 2023

with a non-stationary target. Successively, Wilson et al. (2019) introduces a framework for sequentially solving convex stochastic minimization problems, where the distance between successive minimizers is bounded. The minimization problems are then solved by sequentially applying an optimization algorithm, such as stochastic gradient descent (SGD). In a similar setting, Cao et al. (2019) establishes an upper bound on the regret of a projected SGD algorithm with respect to the drift of the dynamic optima, while Cutler et al. (2021) provides novel non-asymptotic convergence guarantees for stochastic algorithms with iterate averaging.

We study time-varying stochastic optimization in a general statistical setting where we assume we are given a sequence of independent observations $\{X_t\}_{t\in\mathbb{N}}$ with associated densities possessing a parameter that changes through time. In such a framework a problem of interest concerns finding a useful estimator of the time varying parameter at a certain time t - generalizing the classical problem of parameter estimation from the static setting to the time varying parameter setting. Ideally, one would like to find a sequence of estimators that track the time varying parameter through time as closely as possible. We show that, under some assumptions, utilizing the celebrated SGD algorithm (Robbins and Monro 1951) produces a sequence of estimators that will eventually track the time varying parameter - up to a neighborhood - as the number of observations increase.

Established in a general setting that intersects with the frameworks utilized in Cao et al. (2019), Cutler et al. (2021) and Wilson et al. (2019), our results differ from previous work mainly in one aspect: that our objective functions have the specific form of expected log likelihoods, a dissimilarity that will be exploited by utilizing their informational theoretical properties.

The work we present is also linked to the class of score driven models (Creal et al. 2013). Score driven models are a class of observation driven models (here we are using the terminology introduced by Cox et al. (1981)) that update the dynamics of the time varying parameter through the score of the conditional distribution of the observations. Specifically, the same proof technique we utilize to obtain our result can be used to show that a -so called- Newton-score update (Blasques et al. 2015), with the parameter that multiplies the score appropriately chosen, will track the time varying parameter of interest trough time even under possible model misspecificaiton.

A final way to interpret the results we present in this work is as robustness results for a one batch stochastic gradient procedure in the case we are incorrectly assuming that our observations are identically distributed. Indeed, the results show that even if we incorrectly assumed that the true parameter is static (we have IID observations) utilizing a stochastic gradient algorithm with a time dependent single sized batch to optimize the log-likelihood allows us to track the pseudo true time varying parameter up to a neighborhood if it is not moving wildly.

The paper is organised as follows: in Sect. 2 we list and discuss the assumptions of our framework and state the main result. We then present a class of examples given by the exponential family and discuss the performance of SGD with respect to the one observation maximum likelihood estimator at each time. In the third section we provide a detailed proof of our main result.

2 Statement of the main result

Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of independent m-dimensional random vectors defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In the sequel we will write $\mathbb{E}[\cdot]$ for the expected value with respect to the probability measure $\mathbb{P}$, $\|\cdot\|$ for the Euclidean norm in $\mathbb{R}^d$ and $\|\cdot\|_{\mathbb{L}^2(\Omega)}$ for $\mathbb{E}[\|\cdot\|^2]^{\frac{1}{2}}$.

We assume that for any $t \in \mathbb{N}$ the random vector X_t possesses a joint probability density function which depends on the d-dimensional parameter λ_t^*, in symbols $X_t \sim p(\cdot|\lambda_t^*)$. Our aim is to estimate the sequence $\{\lambda_t^*\}_{t\in\mathbb{N}}$ through the observed values $\{X_t\}_{t\in\mathbb{N}}$: To this aim we choose $\lambda_1 \in \mathbb{R}^d$ and utilize the SGD algorithm

$$\lambda_{t+1} := \lambda_t + \alpha \nabla_\lambda \ln p(X_t|\lambda_t), \quad t \in \mathbb{N}. \tag{2.1}$$

Utilizing SGD to attempt to track λ_t^* is motivated by the principle underlying classical maximum likelihood estimation: in fact, under some canonical assumptions we will present below, λ_t^* will be the maximum of the expected log-likelihood $\lambda \to \mathbb{E}[\ln p(X_t|\lambda)]$. Thus, finding a sequence of estimators that track the time varying parameter as closely as possible is connected to finding the maxima of a sequence of expected log-likelihoods, a generalization of the classical static framework. Since we have no direct access to the expected log-likelihoods, but only a singe observation for each time t, we categorize the problem as a time varying *stochastic* optimization problem.

The assumptions we will require to obtain our result are the following.

Assumption 2.1 (Smoothness of the log-likelihood) The function

$$\mathbb{R}^d \ni \lambda \mapsto \ln p(x|\lambda) \tag{2.2}$$

is twice continuously differentiable for all $x \in \mathbb{R}^m$; moreover,

$$\partial_{\lambda_i}\partial_{\lambda_j}\mathbb{E}[\ln p(X_t|\lambda)] = \mathbb{E}\left[\partial_{\lambda_j}\partial_{\lambda_j} \ln p(X_t|\lambda)\right],$$

for all $i, j \in \{1, ..., d\}$ and $t \in \mathbb{N}$.

Assumption 2.2 (Strong convexity) The function in (2.2) is strongly convex uniformly with respect to $x \in \mathbb{R}^m$: i.e., there exists a positive constant ℓ such that for all $x \in \mathbb{R}^m$ the matrix $\mathcal{H}_\lambda[-\ln p(x|\lambda)] - \ell I_d$ is positive semi-definite. Here, $\mathcal{H}_\lambda[-\ln p(x|\lambda)]$ stands for the Hessian matrix of the function in (2.2) while I_d denotes the $d \times d$ identity matrix.

Assumption 2.3 (Lipschitz continuity of the gradient) The function

$$\mathbb{R}^d \ni \lambda \mapsto \nabla_\lambda \ln p(x|\lambda)$$

is globally Lipschitz continuous uniformly with respect to $x \in \mathbb{R}^m$: i.e., there exists a positive constant L such that for all $x \in \mathbb{R}^m$ we have

$$\|\nabla_\lambda \ln p(x|\xi_1) - \nabla_\lambda \ln p(x|\xi_2)\| \le L\|\xi_1 - \xi_2\|, \quad \xi_1, \xi_2 \in \mathbb{R}^d.$$

Assumptions 2.2 and 2.3 are classical in the optimization literature, see for instance Boyd and Vandenberghe (2004) and Bottou et al. (2018); we have utilized the versions of Nesterov (2014). We remark that Assumption 2.2 may seem excessively restrictive at first glance, but we will present in Example 2.9 below a large family of examples where it holds.

Remark 2.4 Assumptions 2.1 and 2.3 imply that

$$\mathtt{I}(\lambda_t^*) \leq dL,$$

where we have denoted $\mathtt{I}(\lambda_t^*) := \mathbb{E}[\|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2]$, i.e. the trace of *Fisher information matrix* of X_t. In fact,

$$\begin{aligned}\mathtt{I}(\lambda_t^*) &= \mathbb{E}[\|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2] = \sum_{j=1}^{d} \mathbb{E}[(\partial_{\lambda_j} \ln p(X_t|\lambda_t^*))^2] = -\sum_{j=1}^{d} \mathbb{E}[\partial^2_{\lambda_j} \ln p(X_t|\lambda_t^*)] \\ &= \sum_{j=1}^{d} \mathbb{E}[\partial^2_{\lambda_j}(-\ln p(X_t|\lambda_t^*))] = \sum_{j=1}^{d} \mathbb{E}[\langle \mathcal{H}_\lambda(-\ln p(X_t|\lambda_t^*))e_j, e_j\rangle] \\ &\leq \sum_{j=1}^{d} \mathbb{E}[\langle L I_d e_j, e_j\rangle] = dL.\end{aligned}$$

We will use Remark 2.4 to bound the quantity $\mathbb{E}[\|\nabla_\lambda \ln p(X_t|\lambda_t)\|^2]$. In the general setting utilized in the optimization literature a bound on $\mathbb{E}[\|\nabla_\lambda \ln p(X_t|\lambda_t)\|^2]$ requires an extra assumption, see Bottou et al. (2018) and the discussion in Nguyen et al. (2018). In our setting we manage to avoid this type of additional assumption thanks to the properties of the Fisher information matrix.

Our last assumption concerns the evolution of the time varying parameter $\{\lambda_t^*\}_{t\in\mathbb{N}}$.

Assumption 2.5 (Lipschitz continuity of the true parameter) There exists a positive constant K such that

$$\|\lambda_{t+1}^* - \lambda_t^*\| \leq K \quad \text{for all } t \in \mathbb{N}.$$

Assumption 2.5 has been used throughout the literature, see for example Simonetto et al. (2020); Cao et al. (2019) and Wilson et al. (2019), since a limitation on the behavior of the sequence of true parameters values must be imposed to be able to track it.

We can now state our main theorem.

Theorem 2.6 *Let Assumptions 2.1, 2.2, 2.3 and 2.5 hold. Then, for* $\alpha \in [\frac{1}{\ell+L}, \frac{1}{L}]$ *running the SGD* (2.1) *we obtain*

$$\limsup_{t\to+\infty} \|\lambda_{t+1} - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} \leq \frac{\varphi(\alpha, L)K + \alpha\sqrt{2dL}}{1 - \varphi(\alpha, L)}, \tag{2.3}$$

where $\varphi(\alpha, L) := \sqrt{1 - 2L\alpha + 2L^2\alpha^2}$. *Moreover, the minimum of the right hand side in* (2.3) *is attained at* $\alpha = \frac{1}{\ell+L}$ *and in this case the last inequality reads*

$$\limsup_{t\to+\infty} \|\lambda_{t+1} - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} \leq \frac{K\sqrt{\ell^2 + L^2} + \sqrt{2dL}}{\ell + L - \sqrt{\ell^2 + L^2}}. \tag{2.4}$$

Remark 2.7 Notice that λ_{t+1} depends on $X_1, X_2, \ldots, X_t$, so as an estimator it is natural to compare it with λ_t^*.

Remark 2.8 In the case of model misspecification, i.e. when the true distribution of the observations is not included in the parametric model $\{p(\cdot|\lambda)\}_{\lambda\in\mathbb{R}^d}$, the same proof technique can be utilized to show that the recursion (2.1) will track the so called *pseudo-true* time varying parameter $\tilde{\lambda}_t$ which is defined as

$$\tilde{\lambda}_t := \arg\max_{\lambda\in\mathbb{R}^d} \mathbb{E}[\ln p(X_t|\lambda)].$$

We recall that the pseudo-true time varying parameter $\tilde{\lambda}_t$ minimizes the Kullback Leiber divergence between the law of the data generating process and the model densities at each time t, see White (1982) and Akaike (1973) for additional details.

The only technical difference in the proof is that Remark 2.4 can't be used since $\mathbb{E}[\|\nabla_\lambda \ln p(X_t|\tilde{\lambda}_t)\|^2]$ is no longer related to the Fisher information matrix of X_t. Thus, an additional assumption is needed to control $E[\|\nabla_\lambda \ln p(X_t|\tilde{\lambda}_t)\|^2]$ but this is standard practice in the optimization literature, see Nguyen et al. (2018) for a discussion on this kind of assumption.

Example 2.9 The exponential family in canonical form provides a class of natural examples where Theorem 2.6 holds. Take as the parameter of interest the natural parameter of a distribution belonging to the exponential family put in canonical form, i.e.

$$p(x|\lambda) = h(x)\exp\{\langle\lambda, T(x)\rangle - A(\lambda)\}, \quad x \in \mathbb{R}^m$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is a non-negative function, $T : \mathbb{R}^m \to \mathbb{R}^d$ is a sufficient statistic and $A : \mathbb{R}^d \to \mathbb{R}$ must be chosen so that $p(x|\lambda)$ integrates to one.

A standard result for exponential families, see for instance Theorem 1.6.3 in Bickel and Doksum (2001), is that A is a convex function of λ; this fact together with identities

$$\nabla_\lambda \ln p(x|\lambda) = T(x) - \nabla_\lambda A(\lambda),$$

and

$$\mathcal{H}_\lambda[-\ln p(x|\lambda)] = -\mathcal{H}_\lambda A(\lambda),$$

implies that one can find, restricting if necessary the range of λ (and hence of $\{\lambda_t^*\}_{t\in\mathbb{N}}$) to a suitable convex compact set Λ, the positive constants l and L required for the validity of Assumptions 2.2-2.3.

Note that the restriction of the range of λ to the convex compact set Λ is carried out by simply modifying (2.1) as

$$\bar{\lambda}_{t+1} := \Pi_\Lambda \left(\bar{\lambda}_t + \alpha \nabla_\lambda \ln p(X_t|\bar{\lambda}_t) \right), \quad t \in \mathbb{N},$$

where Π_Λ denotes the orthogonal projection onto the set Λ. This alternative scheme doesn't affect the validity of Theorem 2.6; in fact, from the contraction property of Π_Λ we get

$$\|\bar{\lambda}_{t+1} - \lambda_t^*\|^2 = \|\Pi_\Lambda(\bar{\lambda}_t + \alpha \nabla_\lambda \ln p(X_t|\bar{\lambda}_t)) - \lambda_t^*\|^2 \leq \|\bar{\lambda}_t + \alpha \nabla_\lambda \ln p(X_t|\bar{\lambda}_t) - \lambda_t^*\|^2,$$

and this corresponds to the first step in the proof of Theorem 2.6 (see Sect. 3 below for more details).

An important question concerning applied settings is whether the estimator λ_t defined in (2.1) performs asymptotically better than the maximum likelihood estimator $\hat{\lambda}_t$ calculated by optimizing the one observation log-likelihood $\ln p(X_t|\lambda_t)$. The following example will showcase that there are indeed cases when utilizing (2.1) is beneficial.

Example 2.10 Referring to Example 2.9 and setting $m = d = 1$ for ease of notation, we consider a sequence of independent observations $\{X_t\}_{t\in\mathbb{N}}$ with

$$X_t \sim p(x|\lambda_t^*) := h(x) \exp\{\lambda_t^* T(x) - A(\lambda_t^*)\}, \quad x \in \mathbb{R}.$$

We assume in addition that $\lambda \mapsto A''(\lambda)$ is continuous and we restrict the parameter space to $\Lambda = [\lambda_m, \lambda_M]$ for suitable real numbers $\lambda_m < \lambda_M$. Observe that Assumptions 2.2 and 2.3 hold in this case with

$$\ell = \min_{\lambda\in\Lambda} A''(\lambda), \quad L = \max_{\lambda\in\Lambda} A''(\lambda).$$

In Theorem 2.6 we obtained an upper bound for the asymptotic mean-square error of λ_t as defined in (2.1). We now want to compare it with the mean-square error of the sufficient statistic $T(X_t)$, which we assume to be unbiased; this means considering the quantity

$$\sqrt{\mathbb{E}[|T(X_t) - \lambda_t^*|^2]} = \sqrt{\mathbb{V}[T(X_t)]} = \sqrt{A''(\lambda_t^*)}, \tag{2.5}$$

where the last equality follows from Theorem 1.6.2 in Bickel and Doksum (2001). Therefore, our estimator λ_t, performs asymptotically better than $T(X_t)$ if

$$\frac{K\sqrt{\ell^2 + L^2} + \sqrt{2L}}{\ell + L - \sqrt{\ell^2 + L^2}} \leq \sqrt{A''(\lambda_t^*)} \quad \text{for all } t \in \mathbb{N}. \tag{2.6}$$

Here, the left hand side corresponds to right hand side in (2.4) with $d = 1$ while the right hand side follows from (2.5). We want this inequality to hold for all possible

values of the sequence $\{\lambda_t^*\}_{t\in\mathbb{N}}$ and this is achieved by taking the infimum of the right hand side of (2.6), i.e., we want

$$\frac{K\sqrt{\ell^2+L^2}+\sqrt{2L}}{\ell+L-\sqrt{\ell^2+L^2}} \leq \sqrt{\ell}. \tag{2.7}$$

A simple investigation of the previous inequality shows that the left hand side increases for small values of ℓ or large values of L; hence, there exist $\bar{\ell}$ and $\bar{L}$ such that for all $\bar{\ell} \leq \ell \leq L \leq \bar{L}$ the asymptotic mean-square error of λ_t is lower than the mean-square error of the sufficient statistic $T(X_t)$. Figures 2 and 3 provide an illustration of this fact. Finally, notice that there are cases when the sufficient statistic of the exponential family is unbiased and coincides with the one observation maximum likelihood estimator, as is the case if we choose as the parameter of interest the variance of a Gaussian.

Example 2.11 A specific member of the exponential family of distributions that leads to pleasing computations is the case of the Gaussian with parameter of interest the mean μ. Setting $m = d = 1$, for ease of notation, the log-likelihood of a Gaussian with mean μ and variance σ^2 is quadratic in μ:

$$\ln p(x|\mu) = -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x-\mu)^2.$$

The second derivative of the negative log-likelihood is

$$-\frac{\partial^2}{\partial\mu^2}\ln p(x|\mu) = \frac{1}{\sigma^2},$$

so it follows that Assumptions 2.2 and 2.3 hold with $\ell = L = \frac{1}{\sigma^2}$. It is also well known that Assumption 2.1 holds in the Gaussian case. Thus, in the specific case of the mean of a Gaussian, Theorem 2.6 tells us that for $\alpha = \frac{1}{\ell+L} = \frac{\sigma^2}{2}$ running the SGD (2.1) we obtain

$$\limsup_{t\to+\infty} \|\mu_{t+1} - \mu_t^*\|_{\mathbb{L}^2(\Omega)} \leq \frac{\sigma(K\sqrt{2}+\sqrt{2d})}{2-\sqrt{2}}. \tag{2.8}$$

In Fig. 1 we simulate the SGD (2.1) given Gaussian observations with constant variance and a time varying mean

Example 2.12 An example outside of the exponential family of distributions is provided, for instance, by a Student-t scale model with exponential link function, i.e.,

$$X = \exp(\lambda)\varepsilon$$

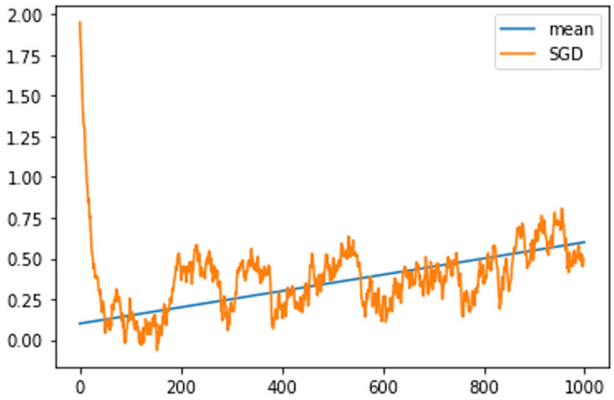


Fig. 1 The trajectory of the SGD (2.1), starting from $\mu_1 = 2$, when Gaussian observations have constant variance ($\sigma^2 = 1$) and a time varying mean that evolves linearly

where ε has a Student-t distribution with degrees of freedom parameter ν. Such a model has, up to additive constants, a log-likelihood given by

$$\ln p(x|\lambda) = -\lambda - \frac{\nu+1}{2} \ln\left(1 + \frac{x^2}{\nu \exp(2\lambda)}\right),$$

the derivative of the log-likelihood with respect to the parameter of interest λ is

$$\frac{\partial}{\partial\lambda} \ln p(x|\lambda) = \frac{(\nu+1)x^2}{\nu \exp(2\lambda) + x^2} - 1.$$

Furthermore, we have that

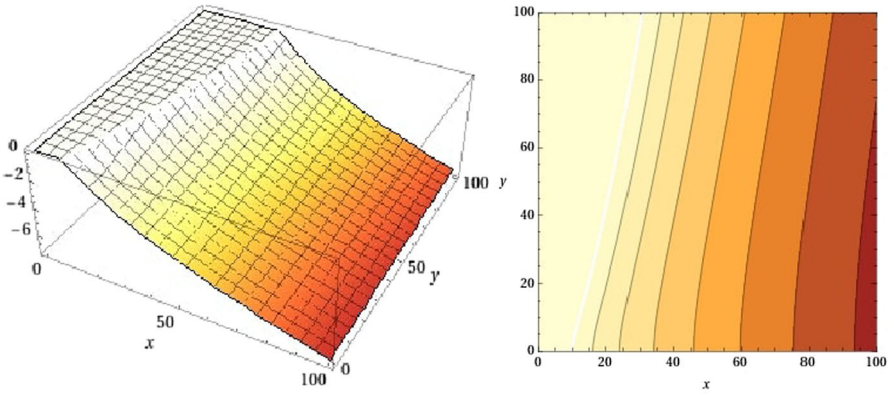
$$-\frac{\partial^2}{\partial\lambda^2} \ln p(x|\lambda) = \frac{2\nu(\nu+1)x^2 \exp(2\lambda)}{(\nu \exp(2\lambda) + x^2)^2},$$

which is strictly positive and uniformly bounded from above and below, so Assumptions 2.2 and 2.3 hold. A model that utilizes a Student-t scale probability distribution with exponential link function in applications is the Beta-t-EGARCH originally proposed by Harvey and Chakravarty (2008), see also Harvey (2013). Other practical settings with a suitable stochastic framework ripe for applications can be found in the actuarial domain, see Maciak et al. (2021).

3 Proof of the main result

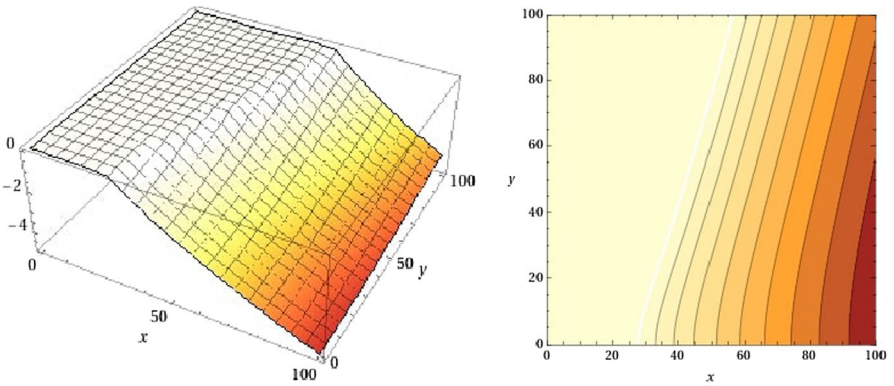
Using (2.1) and expanding the squared Euclidian norm we can write

$$\|\lambda_{t+1} - \lambda_t^*\|^2 = \|\lambda_t - \lambda_t^* + \alpha \nabla_\lambda \ln p(X_t|\lambda_t)\|^2$$



(a) (b)

Fig. 2 Plot of the surface $z = \min\left\{\frac{K\sqrt{\ell^2+L^2}+\sqrt{2}L}{\ell+L-\sqrt{\ell^2+L^2}} - \sqrt{\ell}, 0\right\}$ from (2.7) with $x = l$, $y = L - \ell$ and $K = 1$



(a) (b)

Fig. 3 Plot of the surface $z = \min\left\{\frac{K\sqrt{\ell^2+L^2}+\sqrt{2}L}{\ell+L-\sqrt{\ell^2+L^2}} - \sqrt{\ell}, 0\right\}$ from (2.7) with $x = l$, $y = L - \ell$ and $K = 2$

$$
\begin{aligned}
&= \|\lambda_t - \lambda_t^*\|^2 + 2\alpha\langle\lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t)\rangle + \alpha^2\|\nabla_\lambda \ln p(X_t|\lambda_t)\|^2 \\
&= \|\lambda_t - \lambda_t^*\|^2 + 2\alpha\langle\lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t) - \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle \\
&\quad + 2\alpha\langle\lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle + \alpha^2\|\nabla_\lambda \ln p(X_t|\lambda_t)\|^2 \\
&= \|\lambda_t - \lambda_t^*\|^2 + \mathcal{A}_1 + 2\alpha\langle\lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle + \mathcal{A}_2, \qquad (3.1)
\end{aligned}
$$

where we set

$$\mathcal{A}_1 := 2\alpha\langle\lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t) - \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle$$

and

$$\mathcal{A}_2 := \alpha^2 \|\nabla_\lambda \ln p(X_t|\lambda_t)\|^2.$$

To treat $\mathcal{A}_1$ we employ Theorem 2.1.12 from Nesterov (2014); with $C_1 := \frac{\ell L}{\ell + L}$ and $C_2 = \frac{1}{\ell + L}$ this gives

$$\mathcal{A}_1 \leq -2\alpha C_1 \|\lambda_t - \lambda_t^*\|^2 - 2\alpha C_2 \|\nabla_\lambda \ln p(X_t|\lambda_t) - \nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2; \tag{3.2}$$

moreover, using inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ we get

$$\mathcal{A}_2 \leq 2\alpha^2 \|\nabla_\lambda \ln p(X_t|\lambda_t) - \nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2 + 2\alpha^2 \|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2. \tag{3.3}$$

Combining (3.1) with (3.2) and (3.3) we obtain

$$\begin{aligned}\|\lambda_{t+1} - \lambda_t^*\|^2 \leq & (1 - 2\alpha C_1)\|\lambda_t - \lambda_t^*\|^2 + 2\alpha \langle \lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle \\ & + 2\alpha(\alpha - C_2)\|\nabla_\lambda \ln p(X_t|\lambda_t) - \nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2 \\ & + 2\alpha^2 \|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2.\end{aligned}$$

Imposing that $2\alpha(\alpha - C_2) \geq 0$, or equivalently $\alpha \geq C_2$, we can utilize the Lipschitz continuity of the gradient in the second line above to get

$$\begin{aligned}\|\lambda_{t+1} - \lambda_t^*\|^2 \leq & (1 - 2\alpha C_1 + 2\alpha(\alpha - C_2)L^2)\|\lambda_t - \lambda_t^*\|^2 \\ & + 2\alpha \langle \lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle + 2\alpha^2 \|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2.\end{aligned} \tag{3.4}$$

Notice that according to the definitions of C_1 and C_2 we can write

$$\begin{aligned}1 - 2\alpha C_1 + 2L^2\alpha(\alpha - C_2) &= 1 - 2\alpha(C_1 + L^2 C_2) + 2L^2\alpha^2 \\ &= 1 - 2\alpha\left(\frac{\ell L}{\ell + L} + \frac{L^2}{\ell + L}\right) + 2L^2\alpha^2 \\ &= 1 - 2L\alpha + 2L^2\alpha^2.\end{aligned}$$

therefore, setting $\varphi(\alpha, L) := \sqrt{1 - 2L\alpha + 2L^2\alpha^2}$ inequality (3.4) now reads

$$\begin{aligned}\|\lambda_{t+1} - \lambda_t^*\|^2 \leq & \varphi(\alpha, L)^2 \|\lambda_t - \lambda_t^*\|^2 + 2\alpha \langle \lambda_t - \lambda_t^*, \nabla_\lambda \ln p(X_t|\lambda_t^*)\rangle \\ & + 2\alpha^2 \|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2.\end{aligned}$$

Taking the conditional expectation with respect to the sigma-algebra $\mathcal{F}_{t-1} := \sigma(X_1, ..., X_{t-1})$ of both sides above we obtain

$$\begin{aligned}\mathbb{E}[\|\lambda_{t+1} - \lambda_t^*\|^2 | \mathcal{F}_{t-1}] \leq & \varphi(\alpha, L)^2 \|\lambda_t - \lambda_t^*\|^2 + 2\alpha \langle \lambda_t - \lambda_t^*, \mathbb{E}[\nabla_\lambda \ln p(X_t|\lambda_t^*)|\mathcal{F}_{t-1}]\rangle \\ & + 2\alpha^2 \mathbb{E}[\|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2 | \mathcal{F}_{t-1}]\end{aligned}$$

$$
\begin{aligned}
&= \varphi(\alpha, L)^2 \|\lambda_t - \lambda_t^*\|^2 + 2\alpha \langle \lambda_t - \lambda_t^*, \mathbb{E}[\nabla_\lambda \ln p(X_t|\lambda_t^*)] \rangle \\
&\quad + 2\alpha^2 \mathbb{E}[\|\nabla_\lambda \ln p(X_t|\lambda_t^*)\|^2] \\
&\leq \varphi(\alpha, L)^2 \|\lambda_t - \lambda_t^*\|^2 + 2\alpha^2 dL. \qquad (3.5)
\end{aligned}
$$

Here, we have utilized that

- λ_t is by construction $\mathcal{F}_{t-1}$-measurable for all $t \in \mathbb{N}$;
- the X_t's are independent;
- the expectation of the score is zero, this follows from Assumptions 2.1 and 2.2 and by the fact that λ_t^* is the maximum of the log-likelihood $\lambda \to \ln p(X_t|\lambda)$;
- Remark 2.4.

We now compute the expectation of the first and last members of (3.5) to get

$$
\mathbb{E}[\|\lambda_{t+1} - \lambda_t^*\|^2] \leq \varphi(\alpha, L)^2 \mathbb{E}[\|\lambda_t - \lambda_t^*\|^2] + 2\alpha^2 dL,
$$

which together with inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ gives

$$
\|\lambda_{t+1} - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} \leq \varphi(\alpha, L) \|\lambda_t - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} + \alpha\sqrt{2dL}.
$$

The last step involves using Assumption 2.5 in the previous estimate to obtain

$$
\|\lambda_{t+1} - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} \leq \varphi(\alpha, L) \|\lambda_t - \lambda_{t-1}^*\|_{\mathbb{L}^2(\Omega)} + \varphi(\alpha, L)K + \alpha\sqrt{2dL},
$$

which upon iteration yields

$$
\|\lambda_{t+1} - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} \leq \varphi(\alpha, L)^{t-1} \|\lambda_2 - \lambda_1^*\|_{\mathbb{L}^2(\Omega)} + (\varphi(\alpha, L)K + \alpha\sqrt{2dL}) \frac{1 - \varphi(\alpha, L)^{t-1}}{1 - \varphi(\alpha, L)}.
$$

If $\alpha < \frac{1}{L}$, then $\varphi(\alpha, L) < 1$; we can therefore take the limit as t tends to infinity of both sides to get

$$
\limsup_{t \to +\infty} \|\lambda_{t+1} - \lambda_t^*\|_{\mathbb{L}^2(\Omega)} \leq \frac{\varphi(\alpha, L)K + \alpha\sqrt{2dL}}{1 - \varphi(\alpha, L)};
$$

moreover, the minimum of the right hand side above is attained at $\alpha = \frac{1}{l+L}$ (in view of the constraints needed on α to recover inequality (3.4)).

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Akaike H (1973) Information theory and an extension of the likelihood principle. In: Proceedings of the second international symposium of information theory

Bickel P, Doksum K (2001) Mathematical statistics: basic ideas and selected topics. Number v. 1. Prentice Hall,Hoboken

Blasques F, Koopman SJ, Lucas A (2015) Information-theoretic optimality of observation-driven time series models for continuous responses. Biometrika 102(2):325–343

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60(2):223–311

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge

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

Cox DR, Gudmundsson G, Lindgren G, Bondesson L, Harsaae E, Laake P, Juselius K, Lauritzen SL (1981) Statistical analysis of time series: Some recent developments [with discussion and reply]. Scand J Stat 8(2):93–115

Creal D, Koopman SJ, Lucas A (2013) Generalized autoregressive score models with applications. J Appl Econ 28(5):777–795

Cutler J, Drusvyatskiy D, Harchaoui Z (2021) Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW (eds) Advances in neural information processing systems, vol 34. Curran Associates Inc, pp 11859–11869

Delyon B, Juditsky A (1995) Asymptotical study of parameter tracking algorithms. SIAM J Control Optim 33(1):323–345

Harvey A (2013) Dynamic models for volatility and heavy tails: With applications to financial and economic time series. Time series, Dynamic models for volatility and heavy tails, pp 1–262

Harvey AC, Chakravarty T (2008) Beta-t-(e) garch

Ljung L, Gunnarsson S (1990) Adaptation and tracking in system identification-a survey. Automatica 26(1):7–21

Maciak M, Okhrin O, Pešta M (2021) Infinitely stochastic micro reserving. Insurance 100:30–58

Nesterov Y (2014) Introductory lectures on convex optimization: a basic course, 1st edn. Springer, Berlin

Nguyen L, Nguyen P, Van Dijk M, Richtárik P, Scheinberg K, Takáč M (2018) Sgd and hogwild! convergence without the bounded gradients assumption. In: Krause A, Dy J (eds), 35th international conference on machine learning, ICML 2018, pp 6012–6020. International Machine Learning Society (IMLS). 35th International Conference on Machine Learning, ICML 2018 ; Conference date: 10-07-2018 Through 15-07-2018

Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407

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

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50(1):1–25

Wilson C, Veeravalli VV, Nedić A (2019) Adaptive sequential stochastic optimization. IEEE Trans Automat Control 64(2):496–509

Zhu J, Spall JC (2016) Tracking capability of stochastic gradient algorithm with constant gain. In: 2016 IEEE 55th conference on decision and control (CDC), pp 4522–4527