Sequential Sample Average Majorization–Minimization

Gersende Fort, Florence Forbes, Hien Duy Nguyen

Sequential Sample Average Majorization–Minimization

Gersende Fort[1], Florence Forbes[2] and Hien Duy Nguyen[3]

June 10, 2024

Abstract

Many statistical inference and machine learning methods rely on the ability to optimize an expectation functional, whose explicit form is intractable. The typical method for conducting such optimization is to approximate the expected value problem by a size-N sample average, often referred to as sample average approximation (SAA) or M-estimation. When the solution to the SAA problem cannot be obtained in closed form, the majorization-minimization (MM) algorithm framework constitutes a broad class of incremental optimization solutions, relying on the iterative construction of surrogates, known as majorizers, of the original problem. The ability to solve an SAA problem depends on the availability of all N observations, contemporaneously, which is difficult when N is large or data are observed as a stream. We propose a stochastic MM algorithm that solves the expected value problem via iterative SAA majorizer constructions using sequential subsets of data, which we call *Sequential Sample Average Majorization–Minimization* (`SAM2`). Compared to previous stochastic MM algorithm variants, our method permit an extended definition of majorizers, and does not rely on convexity and smoothness assumptions or make functional restrictions on the class of problems and majorizers. We develop a theory of stochastic convergence for `SAM2`, made possible via the presentation of a novel double array uniform strong law of large numbers. Examples of `SAM2` algorithms are given along with a numerical demonstration of `SAM2` to the quantile regression problem.

[1]CNRS and Institut de Mathématiques de Toulouse, UMR5219; Université de Toulouse; UPS, F-31062 Toulouse Cedex 9, France.
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK 38000 Grenoble, France.
[3]School of Computing, Engineering and Mathematical Sciences, La Trobe Univ., Bundoora 3086, Victoria Australia; Institute of Mathematics for Industry, Kyushu Univ., Nishi Ward, Fukuoka 819-0395, Japan.
All authors thank the ARC grant DP230100905 and the Wombat Inria Associated Team Project. Part of this work (GF) is funded by the Fondation Simone and Cino Del Duca under the project OpSiMorE.

1 Introduction

A common task that often arises when conducting statistical inference or machine learning is to solve the problem of obtaining

$$\mathcal{L}_{\mathsf{F}^\star} \coloneqq \operatorname{argmin}_{\theta \in \mathbb{T}} \mathsf{F}(\theta), \tag{1}$$

where $\mathbb{T}$ is the so-called parameter space, with typical element θ, and $\mathsf{F} : \mathbb{T} \to \mathbb{R}$ is an objective function, defined at each $\theta \in \mathbb{T}$ by

$$\mathsf{F}(\theta) \coloneqq \mathbb{E}\left[\mathsf{f}(\theta, X)\right],$$

the expectation of $\mathsf{f}(\theta, X)$, for some random X in some space $\mathbb{X}$. Typically, the underlying probability measure of X is unknown and thus F cannot be accessed, directly. However, given access to a sample of independent and identically distributed (i.i.d.) replicates of X: $\{X^i, 1 \leq i \leq N\}$, one can approximate problem (1) by the realizable problem of obtaining

$$\mathcal{L}_N \coloneqq \operatorname{argmin}_{\theta \in \mathbb{T}} \frac{1}{N} \sum_{i=1}^{N} \mathsf{f}(\theta, X_i). \tag{2}$$

The approximation of (1) by (2) has appeared throughout the literature under many names, including extremum estimation (Amemiya, 1985; Gourieroux & Monfort, 1995), M-estimation (Serfling, 1980; van der Vaart & Wellner, 2023), minimum contrast estimation (Dacunha-Castelle & Duflo, 1986; Bickel & Doksum, 2015), in statistics, empirical risk minimization (Vapnik, 1998; Vidyasagar, 2003; Shalev-Shwartz & Ben-David, 2014), in machine learning, and sample average approximation (SAA; Bonnans 2019; Shapiro et al. 2021; Cui & Pang 2022) in optimization theory.

In general, (2) is not assumed to be solvable in closed form, and thus iterative optimization routines are required in practice. Given various assumptions regarding the smoothness or convexity of $\mathsf{f}(\cdot, x)$ ($x \in \mathbb{X}$), numerous methods are available for solving (2), including Newton's algorithm and (sub-)gradient descent (see, e.g., Polyak 2021), and derivative-free methods, such as the Nelder–Mead algorithm and pattern search (see, e.g., Audet & Hare 2017), among other techniques. However, such techniques often require the parameter of the problem be vector-valued, which is unnatural in many learning settings, where $\mathbb{T}$ consists of elements that take value in the probability simplex or in the positive definite matrices, as in the problem of estimating finite mixtures of Gaussian distributions, for example (cf. McLachlan & Peel 2000).

EM and MM algorithms The Expectation–Maximization (EM) approach of Dempster et al. (1977) (see also McLachlan & Krishnan 2008) provides a framework for conducting optimization of probabilistic models in potentially non-vector spaces, via the construction of numerical surrogates for f, whose optimization generates a stable sequence of parameter estimates that monotonically improve the objective values. The EM approach can be viewed as a special case of the so-called Majorization–Minimization (MM) algorithms (Hunter & Lange, 2004; Lange, 2013, 2016), which

extend the applicability of the EM algorithm beyond probabilistic models, to more general surrogate constructions for f, to include any function of the form $\mathsf{g} : \mathbb{T} \times \mathbb{X} \times \mathbb{T} \to \mathbb{R}$, fulfilling criterion $\mathsf{f}(\theta, x) \leq \mathsf{g}(\theta, x; \upsilon)$, for any $\theta, \upsilon \in \mathbb{T}$ and $x \in \mathbb{X}$, with equality whenever $\theta = \upsilon$. This permits the inclusion of gradient descent, proximal algorithms and proximal gradient descent, quadratic approximation algorithm (Bohning & Lindsay, 1988), and the convex concave procedure (Yuille & Rangarajan, 2003), among other methods, within the MM algorithm framework. Furthermore, the framework has been extended to permit coordinate and block-wise iterations (Meng & Rubin, 1993; Razaviyayn et al., 2013; Chalvidal et al., 2023), stochastic surrogates (Celeux & Diebolt, 1992; Delyon et al., 1999; Fort & Moulines, 2003), and decentralized computation (Dieuleveut et al., 2021; Cadoni et al., 2016).

Beyond the applicability, inclusivity, and extendability, the MM framework has a strong unified theoretical foundation, with numerous available global convergence and convergence rates results available. Generally applicable results in this direction include the global convergence theory of Vaida (2005), Lange (2013), Byrne (2014), Lange (2016), Lange et al. (2021), and Cui & Pang (2022), as well as the block-wise results of Razaviyayn et al. (2013), and the convergence rates of Mairal (2015), Chouzenoux & Pesquet (2016), and Hong et al. (2017).

Stochastic algorithms As described, the MM algorithms above solve the sample problem (2), for fixed sample size N to produce an estimate for the solution of the expectation problem (1). Via the general SAA theory of Shapiro et al. (2021), it can be shown that the set of solutions of $\mathcal{L}_N$ converges in set deviation to $\mathcal{L}_{\mathsf{F}^\star}$, almost surely, as N gets large. However, the process of solving repeated SAAs as N increases assumes the availability of the entirety of the sequence $\{X^i, 1 \leq i \leq N\}$, for each N, which is infeasible in practice when data are often accessed as a stream, and when computer memory is so as to make storage of the entire sequence of data impossible for large N.

A more feasible scenario is that one has limited access to the data set, via only the subset $\{X^{t,i}, i = 1, \ldots, N_t\}$ at time t, where $\{N_t, t \geq 1\}$ is a sequence of positive integers denoting sample sizes. Such situations arise in the so-called online or iterative algorithms setting, and approaches for solving (1) in such situations are most notably exemplified by the popular stochastic gradient descent algorithms and their variants and refinements, as studied in Shalev-Shwartz & Ben-David (2014), Lan (2020), Lin et al. (2020), Shapiro et al. (2021), among an ever expanding body of literature. As per their deterministic variants, these algorithms require parameters be vector-valued, among other restrictions, and thus are not universal in their applicability. These limitations are similarly shared by stochastic Newton algorithms such as those of Schraudolph et al. (2007), Byrd et al. (2016), and Meyn (2022).

To resolve these limitations, one can again turn to stochastic variants of EM and MM algorithms, which have been comprehensively studied over the years. Examples of such works include the pioneering work of Cappé & Moulines (2009), whose online EM algorithm forms the basis of

the lineage of works including those of Karimi et al. (2019b), Karimi et al. (2019a), Dieuleveut et al. (2021), Karimi & Li (2021), Fort et al. (2021a), and Fort et al. (2021b) regarding the online, stochastic, and mini-batch estimation of probabilistic models. The online EM framework of Cappé & Moulines (2009) was also extended to optimization of general models via the online MM extension of Nguyen et al. (2022). In each of these works, the surrogates of f are required to take a restrictive linearized form, which is required for obtaining theoretical guarantees.

Next, we note the original works of Mairal (2013) and Razaviyayn et al. (2016), who laid the foundation for a convexity-based approaches to theory, where the convergence of the algorithm is obtained via strong convexity and Lipschitz smoothness assumptions on the surrogate of f, with f required to be convex or strongly convex to obtain the convergence rates. Other works following this direction include the works of Liu et al. (2018), Liu et al. (2019), Zhang et al. (2019), Mokhtari & Koppel (2020), Chouzenoux & Fest (2022), Karimi et al. (2022), and Lupu & Necoara (2023).

Finally, we note the approach of Cui & Pang (2022) who exchange the convexity and Lipschitz smoothness of the surrogates with the inclusion of a proximal term. This approach has been further developed in Liu et al. (2022) and Liu & Pang (2023).

Current work and contributions Taking the approach of Cui & Pang (2022) as a starting point, we derive a novel approach for solving problem (1) via a stream of data, using an MM algorithm approach. We call our method *Sequential Sample Average Majorization–Minimization*, or `SAM2` for brevity.

Like the method of Cui & Pang (2022), our approach solves an iterative sequence of optimization problems, characterized by surrogate functions of f, using subsamples $\{X^{t,i}, i = 1, \ldots, N_t\}$ at each iteration t. However, unlike Cui & Pang (2022), our method does not require a proximal term be included as part of the surrogate function, and thus generalizes the approach, since an MM surrogate functions are closed under the addition of a proximal term, as per Cui & Pang (2022), or a Bregman proximal term, as per Lange (2016), Rossignol et al. (2022), and Khanh Hien et al. (2022). As such `SAM2` can be viewed as a direct generalization of the Cui & Pang (2022) method. We develop a theory of the almost-sure convergence of `SAM2`, when the number of iterations tends to infinity. Our theory makes no use of convexity and Lipschitz smoothness assumptions nor restrictions on the form of the surrogate functions and thus provides the least restrictive framework for constructing stochastic MM algorithms, to the best of our knowledge.

In addition to less restrictive assumptions on MM surrogate functions, our approach permits an expanded definition of such surrogates, in that one may take $\upsilon \in \mathbb{U}$ instead of $\upsilon \in \mathbb{T}$, for some arbitrary set $\mathbb{U}$. This permits for a broader interpretation of the MM algorithm framework, following in the spirit of Mairal (2015) who consider the allowance for surrogates to depend on iteration t, and Naderi et al. (2019) who relax the requirement that the surrogate satisfies $\mathsf{f}(\theta, X) = \mathsf{g}(\theta, X; \upsilon)$ for some υ. In particular, the expanded definition allows us to solve problem (1), with f only implicitly defined via the condition that for every $\theta \in \mathbb{T}$, $\mathsf{F}(\theta) \leq \mathbb{E}[\mathsf{g}(\theta, X; \upsilon)]$ for all $\upsilon \in \mathbb{U}$,

and there exists a $\upsilon \in \mathbb{U}$ such that $\mathsf{F}(\theta) = \mathbb{E}[\mathsf{g}(\theta, X; \upsilon)]$.

As a byproduct of proving the global convergence of SAM2, we obtain a uniform strong law of large numbers for double arrays, which is novel, to the best of our knowledge. This strong law provides a useful result with verifiable regularity conditions for functions indexed by compact Euclidean sets, which serves as an alternative to the more generic theory of Ziegler (2001). The result can also be viewed within the context of Andrews (1992), who posit a double array extension of some uniform strong laws, but do not make explicit the required conditions.

Beyond our general convergence results, we also provide detailed investigations of an online mirror descent algorithm and an online proximal-gradient algorithm as examples of special cases of the SAM2 framework. These examples are complemented by numerical results regarding the application of SAM2 to quantile regression, and in particular, the least absolute deviation (LAD) problem. We benchmark our SAM2 algorithm against Stochastic Subgradient descent (SSG), where we demonstrate that the performance of SSG is sensitive to user calibrated step size schedule, which is avoided by SAM2.

The remainder of the manuscript is organized as follows. The SAM2 algorithm framework is described in Section 2. Asymptotic convergence analysis of SAM2 is provided in Section 3. Technical descriptions of online mirror descent and proximal-gradient algorithms as examples are provided in Section 4. Numerical illustrations of SAM2 to the LAD problem appears in Section 5. And finally, proofs and technical results are reported in Section 6.

Notations. Throughout the paper, vectors are column-vectors. $\langle a, b\rangle$ denotes the dot product in $\mathbb{R}^d$ and $\|\cdot\|$ is the associated norm. ∂f is the subdifferential of a function f and ∇f denotes the gradient of a differentiable function f. The theorems, lemmas, corollaries, propositions and examples share the same counter; while the algorithms have a separate counter.

2 The SAM2 algorithm

We consider the following optimization problem on a compact subset $\mathbb{T}$ of $\mathbb{R}^d$

$$\operatorname{argmin}_{\theta\in\mathbb{T}} \mathsf{F}(\theta), \tag{3}$$

when the Majorization–Minimization (MM) framework applies:

A1. There exists a Borel set $\mathbb{U}$ of $\mathbb{R}^{d'}$ and a measurable function $\mathsf{G} : \mathbb{T} \times \mathbb{U} \to \mathbb{R}$ such that

(a) For all $\upsilon \in \mathbb{U}$, $\mathsf{F}(\cdot) \leq \mathsf{G}(\cdot; \upsilon)$ on $\mathbb{T}$.

(b) For all $\theta \in \mathbb{T}$, there exists $\upsilon \in \mathbb{U}$ such that $\mathsf{F}(\theta) = \mathsf{G}(\theta; \upsilon)$.

For all $\theta \in \mathbb{T}$, we define

$$\mathbb{U}[\theta] := \{\upsilon \in \mathbb{U} : \mathsf{F}(\theta) = \mathsf{G}(\theta; \upsilon)\};$$

under A1-b, this set is not empty. We consider the cases when G is an intractable expectation, but stochastic oracles exist.

A2. (a) Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathbb{X}, \mathcal{X})$ be a measurable space. There exists a measurable function $\mathsf{g} : \mathbb{T} \times \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ such that $\mathsf{G}(\theta, \upsilon) := \mathbb{E}[\mathsf{g}(\theta, X; \upsilon)]$; and for all $\theta \in \mathbb{T}$ and $\upsilon \in \mathbb{U}$, $\mathbb{E}[|\mathsf{g}(\theta, X; \upsilon)|] < \infty$.

(b) A stream of random variables $\{X^{t,i}, t \geq 1, 1 \leq i \leq N_t\}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and i.i.d. with the same distribution as X, is available.

Therefore, a stochastic oracle can be defined as follows, computed from N examples:

$$\mathsf{G}(\theta; \upsilon) \approx \frac{1}{N} \sum_{i=1}^{N} \mathsf{g}(\theta, X^i; \upsilon).$$

Example 1. *Consider the minimization of a function* F *on* $\mathbb{T}$*, a compact subset of* $\mathbb{R}^d$*. Assume that* $\mathsf{F}(\theta) := \mathbb{E}[\mathsf{f}(\langle \theta, X \rangle)]$ *where* $\mathsf{f} : \mathbb{R} \to \mathbb{R}$ *is a convex function such that* $\mathbb{E}[|\mathsf{f}(\langle \theta, X \rangle)|] < \infty$ *for all* $\theta \in \widetilde{\mathbb{T}} \supset \mathbb{T}$*.*

By writing $\langle \theta, x \rangle = \langle \upsilon, x \rangle + \sum_{j=1}^{d} (\theta_j - \upsilon_j) x_j$*, it holds since* f *is convex (see (Lange, 2013, Section 8.3))*

$$f(\langle \theta, x \rangle) \leq \mathsf{g}(\theta, x; \upsilon) := \frac{1}{d} \sum_{j=1}^{d} \mathsf{f}\left(\langle \upsilon, x \rangle + d(\theta_j - \upsilon_j) x_j\right), \qquad \upsilon \in \mathbb{T}.$$

We assume that $\widetilde{\mathbb{T}}$ *is large enough so that* $\mathbb{E}[|\mathsf{f}(\langle \upsilon, X \rangle + d(\theta_j - \upsilon_j) X_j)|] < \infty$ *for all* $\theta, \upsilon \in \mathbb{T}$*.*

Set $\mathbb{U} := \mathbb{T}$ *and* $\mathsf{G}(\theta; \upsilon) := \mathbb{E}[\mathsf{g}(\theta, X; \upsilon)]$*. The condition on* $\widetilde{\mathbb{T}}$ *implies that* A*2 is verified. For any* $\upsilon \in \mathbb{T}$*, we have* $\mathsf{g}(\theta, x; \upsilon) - \mathsf{f}(\langle \theta, x \rangle) \geq 0$*, thus implying that* A *1-a holds. Finally* $\mathsf{g}(\theta, x; \theta) = \mathsf{f}(\langle \theta, x \rangle)$ *thus showing that* A *1-b holds and* $\mathbb{U}[\theta] \supseteq \{\theta\}$*. If* f *is strictly convex, then* $\mathbb{U}[\theta] = \{\theta\}$*. This majorizer is particularly useful for converting a* d*-dimensional optimization problem to iterations that solve* d *one-dimensional problems instead. Often, the one-dimensional problems have either computationally efficient or closed form solutions, when the* d*-dimensional problem lacks either.*

Example 2. *Consider the maximization of a positive function* F *on a compact subset* $\mathbb{T}$ *of* $\mathbb{R}^d$*, when* $\mathsf{F}(\theta) := \int \mathsf{f}(\theta, x) \nu(\mathrm{d}x)$ *where* ν *is a* σ*-finite positive measure on* $(\mathbb{X}, \mathcal{X})$ *and* $\mathsf{f} \geq 0$*. Equivalently, we minimize* $-\log \mathsf{F}$*. We assume that* $F < \infty$ *on* $\mathbb{T}$*.*

The MM framework holds for the minimization of $-\log \mathsf{F}$ *since the Jensen's inequality implies that for any* $\tau \in \mathbb{T}$

$$-\log \mathsf{F}(\theta) \leq \mathsf{G}(\theta; \tau) := -\log \mathsf{F}(\tau) - \int \log\left(\frac{\mathsf{f}(\theta, X)}{\mathsf{f}(\tau, X)}\right) \frac{\mathsf{f}(\tau, X)}{\mathsf{F}(\tau)} \nu(\mathrm{d}x).$$

Here, $\mathbb{U} := \mathbb{T}$ *and* A *1-a holds. We have* $\mathsf{G}(\theta; \theta) = -\log \mathsf{F}(\theta)$ *for all* $\theta \in \mathbb{T}$*, so that* A *1-b holds and* $\mathbb{U}[\theta] \supseteq \{\theta\}$*. Set* $\mathsf{g}(\theta, x; \tau) := -\log \mathsf{F}(\tau) - \log\left({}^{\mathsf{f}(\theta,x)}/_{\mathsf{f}(\tau,x)}\right) {}^{\mathsf{f}(\tau,x)}/_{\mathsf{F}(\tau)}$*. Then* A *2 holds under the integrability condition of* $x \mapsto \log\left({}^{\mathsf{f}(\theta,x)}/_{\mathsf{f}(\tau,x)}\right) \mathsf{f}(\tau, x)$ *with respect to* ν*, for all* $\theta, \tau \in \mathbb{T}$*.*

We note that this majorization scheme is exactly that which is used to derive the famous expectation–maximization algorithms of Dempster et al. (1977) (see, also McLachlan & Krishnan, 2008), as per (Lange, 2013, Sec. 9.8) and (Razaviyayn et al., 2013, Sec. 8.5). In that case, f *is the complete data likelihood.*

Example 3. *Consider the minimization of a function* $\mathsf{F} : \mathbb{R}^d \to (-\infty, +\infty]$*, on a compact convex subset* $\mathbb{T}$ *of* $\mathbb{R}^d$*, when* $\mathsf{F}(\theta) := \mathbb{E}[\mathsf{f}(\theta, X)]$*; we assume that* $\mathbb{E}[|\mathsf{f}(\theta, X)|] < \infty$ *for all* $\theta \in \mathbb{T}$ *so that* $\mathbb{T} \subset \mathrm{dom}(\mathsf{F})$*. Let* $\mathfrak{b} : \mathbb{R}^d \to (-\infty, +\infty]$ *be a lower semicontinuous convex function, such that* $\mathbb{T} \subset \mathrm{dom}(\mathfrak{b})$*, continuously differentiable and strictly convex on an open neighborhood* $\mathbb{S}$ *of* $\mathbb{T}$*. Define the Bregman divergence associated to* $\mathfrak{b}$

$$D_{\mathfrak{b}}(\theta; \tau) := \mathfrak{b}(\theta) - \mathfrak{b}(\tau) - \langle \nabla \mathfrak{b}(\tau), \theta - \tau \rangle, \qquad \theta \in \mathbb{R}^d, \tau \in \mathbb{T}. \tag{4}$$

It holds for any $\theta \in \mathbb{R}^d$*,* $x \in \mathbb{X}$*,* $\tau \in \mathbb{T}$ *and* $\gamma > 0$

$$\mathsf{f}(\theta, x) \leq \mathsf{g}(\theta, x; (\tau, \gamma)) := \mathsf{f}(\theta, x) + \frac{1}{\gamma} D_{\mathfrak{b}}(\theta; \tau) + \iota_{\mathbb{T}}(\theta),$$

where $\iota_{\mathbb{T}} : \mathbb{R}^d \to (-\infty, +\infty]$ *is the characteristic function of* $\mathbb{T}$*, defined as* $\iota_{\mathbb{T}}(\theta) = 0$ *if* $\theta \in \mathbb{T}$ *and* $\iota_{\mathbb{T}}(\theta) = +\infty$*, otherwise. Set* $\mathbb{U} := \mathbb{T} \times \mathbb{R}_{>0}$*, and* $\mathsf{G}(\theta; \upsilon) := \mathbb{E}[\mathsf{g}(\theta, X; \upsilon)]$*. Then the condition* A *2 is verified. For any* $\upsilon := (\tau, \gamma) \in \mathbb{U}$*, we have* $\mathsf{g}(\cdot, x; \upsilon) - \mathsf{f}(\cdot, x) \geq 0$ *on* $\mathbb{T}$ *since* $\mathfrak{b}$ *is convex, so that* A *1-a holds. In addition, for all* $\theta \in \mathbb{T}$ *and* $\gamma > 0$*,* $\mathsf{g}(\theta, x; (\theta, \gamma)) = \mathsf{f}(\theta, x)$ *and* $\mathsf{g}(\theta, x; (\tau, \gamma)) \neq \mathsf{f}(\theta, x)$ *when* $\tau \neq \theta$ *since* $\mathfrak{b}$ *is strictly convex; hence,* A *1-b holds and* $\mathbb{U}[\theta] := \{(\theta, \gamma), \gamma > 0\}$*.*

Non uniqueness of the majorizing functions. If $(\theta, \upsilon) \mapsto \mathsf{G}(\theta; \upsilon)$ and $(\theta, x, \upsilon) \mapsto \mathsf{g}(\theta, x; \upsilon)$ satisfy A 1 and A 2, then the functions $(\theta, \upsilon, \tau) \mapsto \mathsf{G}(\theta; \upsilon) + \varphi(D_{\mathfrak{b}}(\varsigma(\theta); \varsigma(\tau)))$ and $(\theta, x, \upsilon, \tau) \mapsto \tilde{\mathsf{g}}(\theta, x; \upsilon, \tau) := \mathsf{g}(\theta, x; \upsilon) + \varphi(D_{\mathfrak{b}}(\varsigma(\theta); \varsigma(\tau)))$ also satisfy A1 and A2 as soon as

(i) $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $\varsigma : \mathbb{T} \to \mathbb{T}$ are measurable functions, and $\varphi(0) = 0$.

(ii) $D_{\mathfrak{b}}(\cdot; \cdot)$ is the Bregman divergence (see (4)) associated to a strictly convex and continuously differentiable function $\mathfrak{b} : \mathbb{S} \to \mathbb{R}$ defined on a neighborhood $\mathbb{S}$ of $\mathbb{T}$.

Note indeed that for all $\theta \in \mathbb{T}$ and $x \in \mathbb{X}$, it holds

$$\begin{cases} \forall (\upsilon, \tau) \in \mathbb{U} \times \mathbb{T} : \mathsf{g}(\theta, x; \upsilon) \leq \tilde{\mathsf{g}}(\theta, x; \upsilon, \tau) \\ \forall \upsilon \in \mathbb{U}[\theta] : \mathsf{g}(\theta, x; \upsilon) = \tilde{\mathsf{g}}(\theta, x; \upsilon, \theta). \end{cases}$$

The introduction of such a Bregman term in the majorization mechanism is the essence of many proximal-type algorithms such as the `Proximal point` algorithm (see e.g. (Polyak, 2021, Section 3.5)) and the `Mirror Descent` algorithm (see Section 4.1) among examples. The motivation is essentially to make the minimization of the majorizing function easier.

Population MM and SAM2 algorithms

When $\operatorname{argmin}_{\mathbb{T}} \mathsf{G}(\cdot; \upsilon)$ is not empty for all υ in the set $\mathbb{U}$ (see e.g. Proposition 11-i for sufficient conditions), then the `Population MM` algorithm given by Algorithm 1 defines a $\mathbb{T}$-valued sequence $\{\theta^t, t \geq 1\}$ satisfying a descent property: $\mathsf{F}(\theta^{t+1}) \leq \mathsf{F}(\theta^t)$ for all $t \geq 0$. We have indeed by A1 and

Algorithm 1 Population MM

Require: an initial value $\theta^0 \in \mathbb{T}$
Ensure: A $\mathbb{T}$-valued sequence $\{\theta^t, t \geq 0\}$.
1: **for** $t = 0, \cdots,$ **do**
2: Choose $\upsilon^t \in \mathbb{U}[\theta^t]$.
3: Compute $\theta^{t+1} \in \underset{\theta \in \mathbb{T}}{\operatorname{Argmin}}\ \mathsf{G}(\theta; \upsilon^t)$.
4: **end for**

by definition of θ^{t+1}: $\mathsf{F}(\theta^{t+1}) \leq \mathsf{G}(\theta^{t+1}; \upsilon^t) \leq \mathsf{G}(\theta^t; \upsilon^t) = \mathsf{F}(\theta^t)$.

When G does not have a closed form expression but stochastic oracles $\mathsf{g}(\theta, X; \upsilon)$ for $\mathsf{G}(\theta; \upsilon)$ are available, a stochastic `Population MM` can be designed. We consider in this paper the *Sequential Sample Average Majorization–Minimization* algorithm (`SAM2`) given by Algorithm 2. Note that under the compactness property of $\mathbb{T}$, the set of minimizers in Equation (5) of Algorithm 2 is not empty as soon as for any $x \in \mathbb{X}$ and $\upsilon \in \mathbb{T}$, the function $\theta \mapsto \mathsf{g}(\theta, x; \upsilon)$ is lower semicontinuous on $\mathbb{T}$.

Algorithm 2 The `SAM2` algorithm

Require: a sequence $\{N_t, t \geq 1\}$ of positive integers, an initial value $\theta^0 \in \mathbb{T}$
Ensure: A $\mathbb{T}$-valued sequence $\{\theta^t, t \geq 0\}$.
1: **for** $t = 0, \cdots,$ **do**
2: Sample a minibatch $\{X^{t+1,i}, i = 1, \cdots, N_{t+1}\}$ of size N_{t+1}.
3: Choose $\upsilon^t \in \mathbb{U}[\theta^t]$.
4: Compute

$$\theta^{t+1} \in \underset{\theta \in \mathbb{T}}{\operatorname{Argmin}} \frac{1}{N_{t+1}} \sum_{i=1}^{N_{t+1}} \mathsf{g}\left(\theta, X^{t+1,i}; \upsilon^t\right). \tag{5}$$

5: **end for**

Remark. In the preceding discussion, we note that $\mathsf{G}(\cdot; \upsilon)$ for $\upsilon \in \mathbb{U}[\theta]$, is a local approximation of F in a neighborhood of $\theta \in \mathbb{T}$ (see A1), whereas the relationship between G and g from A2 is used to estimate G via Monte Carlo simulation. Putting this together, at each iteration of Algorithm 2,

$$\frac{1}{N_{t+1}} \sum_{i=1}^{N_{t+1}} \mathsf{g}\left(\theta, X^{t+1,i}; \upsilon^t\right)$$

is used as an estimate of the approximation $\mathsf{G}(\cdot; \upsilon^t)$ of F. The forthcoming analysis of Algorithm 2 focuses around the control of the numerical and stochastic errors regarding this joint approximation

and estimation process. Such requirement to control both types of errors is not unique to the SAM2 algorithm, as similar analysis is required for gradient-based algorithms when the objective function F is an intractable expectation: gradient-based algorithms operate directly on the sample gradient of F. Unlike such algorithms, however, computation of (5) does not always require information regarding the gradient of F, which is an advantage when such objects are difficult or costly to compute.

3 Asymptotic convergence analysis of SAM2

This section is devoted to the almost-sure convergence of SAM2 when the number of iterations tends to infinity, under the additional assumption that each majorizing function $\mathsf{G}(\cdot;\upsilon)$ possesses a unique minimizer denoted by $\mathsf{T}(\upsilon)$ (see A3-Item b). Regularity conditions on g and the objective function F are also assumed (see A3-Item a,Item c) which imply continuity properties of G and T (see Proposition 11 and Proposition 12). The proof consists in comparing one iteration of SAM2 to one iteration of Population MM and proving that under conditions on the stochastic perturbations, both algorithms have the same limiting set.

We start with discussing the fixed points of Population MM and how this set of points is related to the minimizers of F (see Theorem 4). Then, a general result on the convergence of perturbed iterative scheme is derived (see Theorem 5), whose main assumptions are the existence of a *Lyapunov function* for the iterated map, and the perturbation is vanishing in a sense related to this Lyapunov function; this convergence theorem covers general iterative scheme and is not restricted to the case of the MM one. In the MM framework, the objective function F plays the role of the Lyapunov function and Proposition 7 shows that the vanishing perturbation condition is satisfied as soon as the perturbation is vanishing in a sense related to the majorizing functions G. Combining Theorem 5, Proposition 7 and a condition on the batch size (see A 4) yields the convergence result for SAM2 (see Theorem 9).

The convergence analysis is derived under the following regularity conditions on F and g.

A3. a) For any $x \in \mathbb{X}$, the function $(\theta,\upsilon) \mapsto \mathsf{g}(\theta,x;\upsilon)$ is continuous on $\mathbb{T}\times\mathbb{U}$. In addition, there exists $p>1$ such that $\mathbb{E}\left[\sup_{(\theta,\upsilon)\in\mathbb{T}\times\mathbb{U}} |\mathsf{g}(\theta,X;\upsilon)|^p\right] < \infty$.

b) For all $\upsilon \in \mathbb{U}$, the set $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ has exactly one element, denoted by $\mathsf{T}(\upsilon)$.

c) The function $\theta \mapsto \mathsf{F}(\theta)$ is continuous on $\mathbb{T}$.

Under A3-a, a minimizer θ^{t+1} in (5) exists: SAM2 is well defined. A3-a also implies that $\theta \mapsto \mathsf{G}(\theta;\upsilon)$ is continuous for all $\upsilon \in \mathbb{U}$ (see Proposition 11); since $\mathbb{T}$ is compact, the set $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ has at least one element. Finally, A3-a and A3-b imply that T is a continuous point-to-point map from $\mathbb{U}$ to $\mathbb{T}$ (see Proposition 12).

Example 3 (to follow)**.** *Assume in addition that (i)* F *is a lower semicontinuous convex function on* $\mathbb{R}^d$*; (ii) for all* $x \in \mathbb{X}$*,* $\theta \mapsto \mathsf{f}(\theta, x)$ *is continuous on* $\mathbb{T}$ *and there exists* $p > 1$ *such that* $\mathbb{E}\left[\sup_{\theta\in\mathbb{T}} |\mathsf{f}(\theta, X)|^p\right] < \infty$*; (iii)* γ *is restricted to a compact ball in* $\mathbb{R}_{>0}$*:* $\gamma \in [g_-, g_+]$*.*

The dominated convergence theorem implies that F *is continuous on* $\mathbb{T}$ *so that* A *3-c holds. Since* $\mathfrak{b}$ *is continuous and* γ *is lower bounded from zero, then* A *3-a also holds. In addition, for all* $(\tau, \gamma) \in \mathbb{U}$*,* $\theta \mapsto \mathsf{G}(\theta; (\tau, \gamma))$ *is strictly convex on* $\mathbb{T}$ *since* $\mathfrak{b}$ *is strictly convex and* F *is convex; hence,* $\mathrm{Argmin}_{\theta\in\mathbb{T}}\mathsf{G}(\theta; (\tau, \gamma))$ *is not empty and has exactly one element.*

Theorem 4 provides sufficient conditions for $\mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$ not to be empty and included in the the set $\mathcal{L}$, defined by

$$\mathcal{L} := \{\theta \in \mathbb{T} : \forall \upsilon \in \mathbb{U}[\theta], \forall \tau \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot; \upsilon), \mathsf{F}(\tau) = \mathsf{F}(\theta)\}. \tag{6}$$

$\mathcal{L}$ contains points in $\mathbb{T}$ such that, starting from such a point, one iteration of `Population MM` can not decrease the value of the objective function. Theorem 4 shows that it is included in the subset of $\mathbb{T}$ consisting of points such that, starting from such a point, one iteration of `Population MM` does not decrease the value of the majorizing function whatever it is. Such a property follows from A1, which implies that for all $\theta \in \mathbb{T}$, $\upsilon \in \mathbb{U}[\theta]$ and $\tau \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot; \upsilon)$, it holds

$$\mathsf{F}(\tau) \leq \min_{\mathbb{T}} \mathsf{G}(\cdot; \upsilon) \leq \mathsf{G}(\theta; \upsilon) = \mathsf{F}(\theta). \tag{7}$$

When $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot; \upsilon)$ has exactly one element denoted by $\mathsf{T}(\upsilon)$, then Theorem 4 also shows that

$$\mathcal{L} = \{\theta \in \mathbb{T} : \forall \upsilon \in \mathbb{U}[\theta], \mathsf{F}(\mathsf{T}(\upsilon)) = \mathsf{F}(\theta) = \min_{\mathbb{T}}\mathsf{G}(\cdot; \upsilon)\} = \{\theta \in \mathbb{T} : \forall \upsilon \in \mathbb{U}[\theta], \mathsf{T}(\upsilon) = \theta\}, \tag{8}$$

so that $\mathcal{L}$ is the set of the fixed points of `Population MM` (see Algorithm 1). The proof of Theorem 4 is given in Section 6.2.

Theorem 4. *i) Assume* A *3-a. For all* $\upsilon \in \mathbb{U}$*, the set* $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot; \upsilon)$ *is not empty.*

ii) Assume A *1,* A *3-a and* A *3-c. The set* $\mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$ *is not empty and* $\mathrm{Argmin}_{\mathbb{T}}\mathsf{F} \subset \mathcal{L} \subset \{\theta \in \mathbb{T} : \theta \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot; \upsilon)$ *for all* $\upsilon \in \mathbb{U}[\theta]\}$*.*

iii) Assume A *1 and* A *3. Then the equalities* (8) *hold.*

Example 3 (to follow)**.** *Under the stated assumptions,* A *1 and* A *3 are verified. Let us identify the set* $\mathcal{L}$*.*

Let $\theta \in \mathbb{T}$ *and* $\upsilon = (\theta, \gamma) \in \mathbb{U}[\theta]$*: for* $\upsilon := (\theta, \gamma)$*, since* $\theta \mapsto \mathsf{G}(\theta; \upsilon)$ *is a lower semicontinuous proper convex function then it holds (see e.g. (Bauschke & Combettes, 2011, Theorem 16.3)):* $0 \in \partial\left[\mathsf{G}(\cdot; \upsilon)\right](\mathsf{T}(\upsilon))$*. By (Bauschke & Combettes, 2011, Corollary 16.48 and Proposition 16.6), we have*

$$\partial\left[\mathsf{G}(\cdot; \upsilon)\right](\mathsf{T}(\upsilon)) = \gamma\,\partial\left[\mathsf{F} + \iota_{\mathbb{T}}\right](\mathsf{T}(\upsilon)) + \nabla\mathfrak{b}(\mathsf{T}(\upsilon)) - \nabla\mathfrak{b}(\theta).$$

Therefore, the condition $\mathsf{T}(\upsilon) = \theta$ *is equivalent to* $0 \in \partial\left[\mathsf{F} + \iota_{\mathbb{T}}\right](\theta)$*: the set* $\mathcal{L}$ *is the set of the minimizers of* F *on* $\mathbb{T}$*. Note also that we have* $\mathcal{L} = \{\theta \in \mathbb{T} : \exists \upsilon \in \mathbb{U}[\theta], \mathsf{T}(\upsilon) = \theta\}$*.*

The proof of the asymptotic convergence consists in observing that `SAM2` is a stochastic perturbation of the `Population MM` iterative scheme $\tau^{t+1} = \mathsf{T}(\upsilon^t)$ for $\upsilon^t \in \mathbb{U}[\tau^t]$. While the sequence $\{\tau^t, t \geq 0\}$ possesses a Lyapunov function i.e. $\mathsf{F}(\tau^{t+1}) \leq \mathsf{F}(\tau^t)$ for all $t \geq 0$, this is no more the case for the `SAM2` sequence $\{\theta^t, t \geq 0\}$. Indeed, under A3-b, θ^{t+1} is the minimizer of an approximation of the majorizing function $\mathsf{G}(\cdot; \upsilon^t)$ so that in general, $\mathsf{F}(\theta^{t+1}) \neq \mathsf{F}(\mathsf{T}(\upsilon^t))$. Nevertheless, the following result shows that as soon as the error $|\mathsf{F}(\theta^{t+1}) - \mathsf{F}(\mathsf{T}(\upsilon^t))|$ vanishes when $t \to +\infty$, the sequence $\{\theta^t, t \geq 0\}$ inherits the same limiting behavior as the `Population MM` sequence. The proof of Theorem 5 is given in Section 6.3; it is adapted from (Fort & Moulines, 2003, Proposition 9) which addresses the case $\mathbb{U}[\theta] = \{\theta\}$ for all $\theta \in \mathbb{T}$.

Theorem 5. *Let* $\mathbb{T}$ *be a compact subset of* $\mathbb{R}^d$*,* $\mathbb{U}$ *be a subset of* $\mathbb{R}^{d'}$ *and let* $\mathcal{L} \subseteq \mathbb{T}$ *be a set such that* $\mathcal{L} \cap \mathbb{T}$ *is compact. Let* $\mathsf{T} : \mathbb{U} \to \mathbb{T}$ *and* $\mathsf{F} : \mathbb{T} \to \mathbb{R}$ *be a continuous function such that*

(H-i) for all $\theta \in \mathbb{T}$*, there exists* $\mathbb{U}[\theta] \subseteq \mathbb{U}$ *such that* $\mathsf{F}(\mathsf{T}(\upsilon)) \leq \mathsf{F}(\theta)$ *for all* $\upsilon \in \mathbb{U}[\theta]$*.*

(H-ii) for any compact subset $\mathcal{K}$ *in* $\mathbb{T} \setminus \mathcal{L}$*,* $\inf_{\theta \in \mathcal{K}} \inf_{\upsilon \in \mathbb{U}[\theta]} \left(\mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\upsilon))\right) > 0$*.*

Let $\{(\theta^t, \upsilon^t), t \geq 0\}$ *be a* $\mathbb{T} \times \mathbb{U}$*-valued sequence such that* $\upsilon^t \in \mathbb{U}[\theta^t]$ *for all* $t \geq 0$ *and*

(H-iii) $\lim_t |\mathsf{F}(\mathsf{T}(\upsilon^t)) - \mathsf{F}(\theta^{t+1})| = 0$*.*

Then the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ *converges to a connected component of* $\mathsf{F}(\mathcal{L} \cap \mathbb{T})$*. If* $\mathsf{F}(\mathcal{L} \cap \mathbb{T})$ *has an empty interior, the sequence* $\{\mathsf{F}(\theta^t), t \geq 0\}$ *converges to* $\mathsf{F}^\star$ *and the sequence* $\{\theta^t, t \geq 0\}$ *converges to the set* $\mathbb{T} \cap \mathcal{L}_{\mathsf{F}^\star}$ *where* $\mathcal{L}_{\mathsf{F}^\star} := \{\theta \in \mathcal{L} : \mathsf{F}(\theta) = \mathsf{F}^\star\}$*.*

Let us start with discussing H-i and H-ii on two examples.

Example 6. *Assume A 1, A 2 and A 3, and* $\mathbb{U}[\theta] = \{\theta\}$ *for all* $\theta \in \mathbb{T}$*. We prove the conditions H-i and H-ii are verified with* $\mathcal{L} := \{\theta \in \mathbb{T} : \mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\theta)) = 0\} = \{\theta \in \mathbb{T} : \mathsf{T}(\theta) = \theta\}$*.*

A 3-b and A 3-c imply that T *is a point-to-point map and* F *is a continuous function. Since* F *and* T *are continuous on* $\mathbb{T}$ *(see A 3-c and Proposition 12), the set* $\mathcal{L}$ *is closed which implies that* $\mathcal{L} \cap \mathbb{T}$ *is compact. By A 1 (see also (7)),* $\mathsf{F}(\mathsf{T}(\theta)) \leq \mathsf{F}(\theta)$ *for all* $\theta \in \mathbb{T}$*, so that H-i holds. By definition of* $\mathcal{L}$*,* $\mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\theta)) > 0$ *for all* $\theta \in \mathbb{T} \setminus \mathcal{L}$*; since* $\theta \mapsto \mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\theta))$ *is continuous on* $\mathbb{T}$*, the condition H-ii holds.*

Example 3 (to follow)**.** *We prove the conditions H-i and H-ii are verified.*

H-i follows from A 1, which holds true for this example. For H-ii, observe that by assumption $\mathbb{U}[\theta] = \{(\theta, \gamma), \gamma \in [g_-, g_+]\}$ *and remember that* F *and* T *are continuous functions on* $\mathbb{T}$ *and* $\mathbb{U}$ *respectively (see Proposition 12). If* $\theta \in \mathbb{T} \setminus \mathcal{L}$ *then* θ *is not a minimizer of* F *on* $\mathbb{T}$*, and* $\theta \neq \mathsf{T}(\theta, \gamma)$

for **all** $(\theta,\gamma) \in \mathbb{U}[\theta]$. *Then,* $(\theta,\gamma) \mapsto \mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\theta,\gamma))$ *is a positive continuous function on* $\mathcal{K} \times [g_-, g_+]$ *for any compact set* $\mathcal{K} \subset \mathbb{T} \setminus \mathcal{L}$. *This establishes H-ii.*

When γ *is not assumed lower bounded away from zero, the condition H-ii may not hold. As a counter-example, consider the minimization of the absolute value function* $\mathsf{F}(\theta) = |\theta|$ *on the compact subset* $\mathbb{T} := [-a, a]$ *of* $\mathbb{R}$*; choose* $\mathfrak{b}(\tau) := \tau^2$. *Then* $\mathsf{T}(\theta,\gamma) = \text{sign}(\theta)(|\theta| - \gamma)_+$ *which implies that* $\mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\theta,\gamma)) = |\theta| \wedge \gamma$. *Therefore,* $\inf_{\theta\in\mathcal{K}} \inf_{\gamma>0} \mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\theta,\gamma)) = 0$ *for any compact subset* $\mathcal{K}$ *of* $[-a, 0) \cup (0, a]$.

Checking H-iii is specific to each perturbation of `Population MM` since it relies on the definition of θ^{t+1} given θ^t and υ^t. The following result, whose proof is given in Section 6.4, provides conditions implying that if θ' is such that $\mathsf{G}(\theta';\upsilon) - \min_{\mathbb{T}} \mathsf{G}(\cdot;\upsilon)$ is small, then $|\mathsf{F}(\theta') - \mathsf{F}(\mathsf{T}(\upsilon))|$ is small.

Proposition 7. *Assume* A *1 and* A *3. Assume also that for any* $\delta > 0$, *there exists* $\tilde{\eta}_\delta > 0$ *such that*

$$(\theta,\theta') \in \mathbb{T}\times\mathbb{T}, \upsilon \in \mathbb{U}[\theta] : \|\theta' - \mathsf{T}(\upsilon)\| \geq \delta \Longrightarrow \Big(\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon)\Big) \geq \tilde{\eta}_\delta. \tag{9}$$

For any $\epsilon > 0$, *there exists* $\alpha_\epsilon > 0$ *such that for any* $\theta, \theta' \in \mathbb{T}$ *and for all* $\upsilon \in \mathbb{U}[\theta]$, *it holds*

$$|\mathsf{F}(\theta') - \mathsf{F}(\mathsf{T}(\upsilon))| \leq \epsilon + \alpha_\epsilon \sup_{\mathbb{T}} |\mathsf{F}| \ \text{Diam}(\mathbb{T}) \ \Big(\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon)\Big),$$

where Diam($\mathbb{T}$) *denotes the diameter of the compact set* $\mathbb{T}$.

Example 8 (Example 6 to follow)**.** *Let us check the condition* (9). *Since* $\mathbb{U}[\theta] = \{\theta\}$, *we have* $\theta' - \mathsf{T}(\upsilon) = \theta' - \mathsf{T}(\theta)$ *and* $\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon) = \mathsf{G}(\theta';\theta) - \mathsf{G}(\mathsf{T}(\theta);\theta)$. *By Proposition 12,* T *is a continuous function so that the set* $\mathcal{K}_\delta := \{(\theta,\theta') \in \mathbb{T}\times\mathbb{T} : \|\theta' - \mathsf{T}(\theta)\| \geq \delta\}$ *is a compact subset of* $\mathbb{T}\times\mathbb{T}$. $\mathsf{G}(\mathsf{T}(\theta);\theta) = \min_{\mathbb{T}} \mathsf{G}(\cdot;\theta)$ *and the minimizer is unique by* A *3-b; therefore,* $\mathsf{G}(\theta';\theta) - \mathsf{G}(\mathsf{T}(\theta);\theta) > 0$ *for all* $(\theta',\theta) \in \mathcal{K}_\delta$. *Finally,* G *and* T *are continuous (see Proposition 11 and Proposition 12). Hence* $\tilde{\eta}_\delta$ *exists.*

Example 3 (to follow)**.** *The proof of* (9) *is on the same lines as in the Example 8. Set* $\mathcal{K}_\delta := \{(\theta,\theta',\gamma) \in \mathbb{T}\times\mathbb{T}\times[g_-,g_+] : \|\theta' - \mathsf{T}(\theta,\gamma)\| \geq \delta\}$. *Under the stated assumptions,* T *is continuous on* $\mathbb{T}\times[g_-,g_+]$ *so* $\mathcal{K}_\delta$ *is compact. In addition,* $\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon) > 0$ *for all* $(\theta',\upsilon) \in \mathcal{K}_\delta$. *Finally,* G *is a continuous function. Hence* $\tilde{\eta}_\delta$ *exists.*

We are now ready to apply Theorem 5 and provide a result on the asymptotic convergence of `SAM2`. Theorem 9 establishes the almost-sure convergence of the `SAM2` sequence to the set $\mathcal{L} = \{\theta \in \mathbb{T} : \mathsf{T}(\upsilon) = \theta \text{ for all } \upsilon \in \mathbb{U}[\theta]\}$. In order to satisfy the condition H-iii of Theorem 5, we apply Proposition 7 and provide sufficient conditions for the property $\lim_t \{\mathsf{G}(\theta^{t+1};\upsilon^t) - \mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t)\} = 0$ to hold almost-surely. Upon noting that

$$0 \leq \mathsf{G}(\theta^{t+1};\upsilon^t) - \mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t) \leq 2 \sup_{(\theta,\upsilon)\in\mathbb{T}\times\mathbb{U}} \left| \frac{1}{N_{t+1}} \sum_{i=1}^{N_{t+1}} \mathsf{g}(\theta, X^{t+1,i};\upsilon) - \mathsf{G}(\theta;\upsilon) \right|,$$

(see Section 6.6), we use a uniform strong law of large numbers (see Section 6.5) and assume that the batch size N_t increases rapidly enough.

A4. The size $\{N_t, t \geq 1\}$ of the mini-batch $\#t$ satisfies $\sum_{t\geq 1} N_t^{-((p-1)\wedge(p/2))} < \infty$, where p is given by A3-a.

When $p \in (1, 2]$, the batch size N_t can be set to $N_t \propto t^{1/(p-1)}(\ln t)^{(1+\beta)/(p-1)}$ for $\beta > 0$. When $p \geq 2$, the batch size N_t can be set to $N_t \propto t^{2/p}(\ln t)^{(1+\beta)2/p}$ for $\beta > 0$. A large value of p yields a slow increase of the batch size N_t.

Theorem 9. *Assume A 2, A 3-a, A 3-b, A 4 and $\mathbb{U}$ is compact. Let $\{\theta^t, t \geq 0\}$ be the output of the SAM2 algorithm. Then with probability one*

i) $\lim_t \{\mathsf{G}(\theta^{t+1}; \upsilon^t) - \mathsf{G}(\mathsf{T}(\upsilon^t); \upsilon^t)\} = 0$.

Set

$$\mathcal{L}^+ := \{\theta \in \mathbb{T} : \exists \upsilon \in \mathbb{U}[\theta], \mathsf{T}(\upsilon) = \theta\}.$$

If in addition A 1 and A 3-c hold, then with probability one,

ii) $\{\mathsf{F}(\theta^t), t \geq 0\}$ *converges to a connected component of* $\mathsf{F}(\mathcal{L}^+)$.

iii) If $\mathsf{F}(\mathcal{L}^+)$ *has an empty interior, then* $\{\mathsf{F}(\theta^t), t \geq 0\}$ *converges to* $\mathsf{F}^\star$ *and* $\{\theta^t, t \geq 0\}$ *converges to* $\{\theta \in \mathcal{L}^+ : \mathsf{F}(\theta) = \mathsf{F}^\star\}$.

The proof of Theorem 9 is given in Section 6.6. When $\mathbb{U}[\theta] = \{\theta\}$ for all $\theta \in \mathbb{T}$, then $\mathcal{L}^+ = \mathcal{L}$ where $\mathcal{L}$ is given by (7). Note also that for Example 3, $\mathcal{L}^+ = \mathcal{L}$ even though $\mathbb{U}[\theta]$ is not reduced to a singleton.

When the majorizing function $\mathsf{G}(\cdot; \upsilon^t)$ can be exactly evaluated at each iteration, it is known that the convergence of the objective function along the path $\{\tau^t, t \geq 0\}$ of the `Population MM` algorithm follows from the monotonicity of $t \mapsto \mathsf{F}(\tau^t)$ (see e.g. (Lange et al., 2021, Proposition 2.1)). When only oracles of the majorizing function are available, the monotonicity property along the `SAM2` sequence does not hold anymore; nevertheless, the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ may still converge, as shown by Theorem 9.

Since there is no guarantee that for the `SAM2` sequence, $\theta^{t+1} \in \{\theta \in \mathbb{T} : \mathsf{F}(\theta) \leq \mathsf{F}(\theta^t)\}$, the stability of the `SAM2` sequence can not follow from the compactness of the level sets of the descent function F. In this paper, the stability is ensured by forcing a compact-valued sequence (see (5) and the compactness assumption on $\mathbb{T}$). Nevertheless, a self-stabilization by projections on growing compact sets could be explored, by adapting (Fort & Moulines, 2003, Propositions 10 and 11). Such an extension is out of the scope of this paper.

Iterative algorithms having a descent function converge to the set of the so-called *no-progress* points (see e.g. (Lange et al., 2021, Proposition 2.2.)). Under A1 and the assumption $\mathbb{U}[\theta] = \{\theta\}$, we have $\mathsf{F}(\tau) \leq \mathsf{F}(\theta)$ for all $\tau \in \mathrm{Argmin}_{\mathbb{T}} \mathsf{G}(\cdot; \theta)$ and the inequality is strict when $\theta \notin \mathrm{Argmin}_{\mathbb{T}} \mathsf{G}(\cdot; \theta)$.

Hence, the set of the no-progress points is the set $\mathcal{L}$ given by (6). Theorem 9 provides sufficient conditions for the `SAM2` sequence to converge to $\mathcal{L}$ (observe that $\mathcal{L} = \mathcal{L}^+$ when $\mathbb{U}[\theta] = \{\theta\}$ and $\text{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\theta)$ has exactly one element).

4 SAM2 examples

4.1 An Online Mirror Descent Algorithm

The goal is to minimize $\mathsf{F} : \mathbb{R}^d \to \mathbb{R}$ on a compact convex set $\mathbb{T}$, via a Mirror Descent approach (see Nemirovskii & Yudin (1983), see also the overview Bubeck (2015)). Consider the case when

MD 1. (a) F is continuously differentiable in an open neighborhood $\mathbb{S}$ of $\mathbb{T}$. There exists a measurable function $\mathsf{h} : \mathbb{T} \times \mathbb{X} \to \mathbb{R}^d$ such that $\nabla\mathsf{F}(\theta) = \mathbb{E}\left[\mathsf{h}(\theta, X)\right]$ where $\mathbb{E}\left[\|\mathsf{h}(\theta, X)\|\right] < \infty$ for all $\theta \in \mathbb{T}$. No exact computation of the expectation is available.

(b) A stream of independent random variables with the same distribution as X is available, and $\mathsf{h}(\theta, x)$ can be computed for all $(\theta, x) \in \mathbb{T} \times \mathbb{X}$.

Choose a function ψ satisfying

MD 2. $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ is a lower semicontinuous convex function, $\mathbb{T} \subset \text{dom}(\psi)$, and it is a continuously differentiable strictly convex function on $\mathbb{S}$. The function $\mathfrak{b} := \psi - \mathsf{F}$ is strictly convex.

Define the Bregman divergences associated to ψ and $\mathfrak{b}$

$$D_\psi\left(\theta;\tau\right) := \psi(\theta) - \psi(\tau) - \left\langle \nabla\psi(\tau), \theta - \tau\right\rangle, \qquad D_{\mathfrak{b}}\left(\theta;\tau\right) := \mathfrak{b}(\theta) - \mathfrak{b}(\tau) - \left\langle \nabla\mathfrak{b}(\tau), \theta - \tau\right\rangle.$$

The `Online Mirror Descent` algorithm is given by Algorithm 3. The update mechanism (10) is

Algorithm 3 The `Online Mirror Descent` algorithm

Require: a sequence $\{N_t, t \geq 1\}$ of positive integers, an initial value $\theta^0 \in \mathbb{T}$
Ensure: A $\mathbb{T}$-valued sequence $\{\theta^t, t \geq 0\}$.

1: **for** $t = 0, \cdots,$ **do**
2: Sample a minibatch $\left\{X^{t+1,i}, i = 1, \cdots, N_{t+1}\right\}$ of size N_{t+1}.
3: Set $\mathcal{H}^{t+1} := N_{t+1}^{-1} \sum_{i=1}^{N_{t+1}} \mathsf{h}\left(\theta^t, X^{t+1,i}\right)$
4: Compute

$$\theta^{t+1} \in \underset{\theta\in\mathbb{T}}{\text{Argmin}} \left\langle \mathcal{H}^{t+1}, \theta\right\rangle + D_\psi\left(\theta;\theta^t\right). \tag{10}$$

5: **end for**

equivalent to

$$\theta^{t+1} = \text{Argmin}_{\theta\in\mathbb{T}}\left(\psi(\theta) - \psi(\theta^t) - \left\langle \nabla\psi(\theta^t) - \mathcal{H}^{t+1}, \theta - \theta^t\right\rangle\right)$$

thus showing that it is `Mirror Descent` applied with a constant step size γ (set to one here without loss of generality, since ψ can be replaced with ψ/γ). Note also that the minimizer exists and is unique since ψ is strictly convex on $\mathbb{T}$.

Online Mirror Descent is a `SAM2` algorithm. Following the same lines as in (Beck & Teboulle, 2003, section 3) (see also (Bubeck, 2015, Section 4.2.), (Lange et al., 2021, Section 4.2.2)), we prove that under MD1 and MD2, Algorithm 3 is a `SAM2` algorithm. Set $\mathbb{U} := \mathbb{T}$ and define

$$\mathsf{G}(\theta; \upsilon) := \mathsf{F}(\upsilon) + \psi(\theta) - \psi(\upsilon) - \langle \nabla\psi(\upsilon) - \nabla\mathsf{F}(\upsilon), \theta - \upsilon\rangle + \iota_{\mathbb{T}}(\theta), \qquad \theta \in \mathbb{R}^d, \upsilon \in \mathbb{T}; \tag{11}$$

$\iota_{\mathbb{T}}$ is the characteristic function of $\mathbb{T}$. For $\theta, \upsilon \in \mathbb{T}$, we have $\mathsf{G}(\theta; \upsilon) = \mathsf{F}(\theta) + D_{\mathfrak{b}}(\theta; \upsilon) + \iota_{\mathbb{T}}(\theta)$. Therefore, $\mathsf{G}(\theta, \upsilon) \geq \mathsf{F}(\theta)$ for all $\theta, \upsilon \in \mathbb{T}$ since $D_{\mathfrak{b}}(\theta; \upsilon) \geq 0$ under MD2. In addition, $\mathsf{G}(\theta; \theta) = \mathsf{F}(\theta)$ for all $\theta \in \mathbb{T}$ since $D_{\mathfrak{b}}(\theta; \theta) = 0$, and since $\mathfrak{b}$ is strictly convex, we have $D_{\mathfrak{b}}(\theta; \upsilon) > 0$ for $\theta \neq \upsilon$. Hence, $\mathbb{U}[\theta] = \{\theta\}$. From (11), we have $\mathsf{G}(\theta, \upsilon) = \mathbb{E}[\mathsf{g}(\theta, X; \upsilon)]$ where g is given by

$$\mathsf{g}(\theta, x; \upsilon) := \langle \mathsf{h}(\upsilon, x), \theta\rangle + D_{\psi}(\theta; \upsilon) + \mathcal{C}(\upsilon) + \iota_{\mathbb{T}}(\theta);$$

$\mathcal{C}(\upsilon)$ does not depend on θ. Hence, it is readily seen that solving (5) is equivalent to solving (10): Online Mirror Descent is a `SAM2` algorithm.

The limiting set. Under MD1, MD2 and

MD3. for all $x \in \mathbb{X}$, $\theta \mapsto \mathsf{h}(\theta, x)$ is continuous on $\mathbb{T}$ and $\mathbb{E}[\sup_{\theta\in\mathbb{T}} \|\mathsf{h}(\theta, X)\|^p] < \infty$ for some $p > 1$

the conditions A1, A2 and A3 are satisfied. Since $\mathbb{U}[\theta] = \{\theta\}$, then $\mathcal{L}^+ = \mathcal{L}$; let us identify the set $\mathcal{L}$ given by (6). Since $\theta' \mapsto \mathsf{G}(\theta'; \theta)$ is a lower semicontinuous convex function with domain $\mathbb{T}$, $\mathsf{T}(\theta)$ solves (see e.g. (Bauschke & Combettes, 2011, Theorem 16.3))

$$0 \in \partial\left[\psi(\cdot) - \langle \nabla\psi(\theta) - \nabla\mathsf{F}(\theta), \cdot\rangle + \iota_{\mathbb{T}}(\cdot)\right](\mathsf{T}(\theta)).$$

Hence, $\mathsf{T}(\theta) = \theta$ is equivalent to $0 \in \nabla\mathsf{F}(\theta) + \partial\iota_{\mathbb{T}}(\mathsf{T}(\theta))$ by (Bauschke & Combettes, 2011, Corollary 16.48) . Hence $\mathcal{L}$ is the set of the minimizers of F on $\mathbb{T}$.

For the sequence $\{\theta^t, t \geq 0\}$ given by Algorithm 3, Theorem 9 shows that, as soon as the size of the mini batches is chosen so that A4 holds: the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ converges to $\min_{\mathbb{T}} \mathsf{F}$ and the sequence $\{\theta^t, t \geq 0\}$ converges to the set of the minimizers.

When the majorizing function G relies on a Bregman divergence, as it is the case in this example, the convergence analyses usually adopt a different approach which uses the *three-point inequality* by Chen & Teboulle (1993): the function that plays the role of a descent function is not F but it is a Bregman distance between the current iterate and a minimizer. In that approach, ϵ-approximate stationarity is generally discussed, most often along an averaged path (see e.g. Nemirovski et al. (2009); Lan et al. (2012); Nedić & Lee (2014); Zhang & He (2018);

Lan (2020); Dragomir et al. (2021); D'Orazio et al. (2023)), from which asymptotic convergence can be obtained (see e.g. (Beck & Teboulle, 2003, Theorem 4.1) in the deterministic case, or Lei & Zhou (2020) for convergence in expectation of `Online Mirror Descent`). While our approach addresses the almost-sure convergence under a constant step size, but at the price of a mini-batch size N_t increasing at each iteration, the other approach considers the case $N_t = 1$ but the step size is iteration-varying and its choice is crucial.

4.2 An Online Proximal-Gradient algorithm

The objective is to minimize a composite function $\mathsf{F} := \mathsf{F}_s + \mathsf{F}_c$ on a compact convex subset $\mathbb{T}$ of $\mathbb{R}^d$, where

PG1. (a) the function $\mathsf{F}_s : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable with L-Lipschitz gradient. There exists a measurable function $\mathsf{h} : \mathbb{T} \times \mathbb{X} \to \mathbb{R}^d$ such that $\nabla \mathsf{F}_s(\theta) = \mathbb{E}\left[\mathsf{h}(\theta, X)\right]$ where $\mathbb{E}\left[\|\mathsf{h}(\theta, X)\|\right] < \infty$ for all $\theta \in \mathbb{T}$. No exact computation of the expectation is available.

(b) A stream of independent random variables with the same distribution as X is available, and $\mathsf{h}(\theta, x)$ can be computed for all $(\theta, x) \in \mathbb{T} \times \mathbb{X}$.

(c) the function $\mathsf{F}_c : \mathbb{R}^d \to (-\infty, +\infty]$ is a continuous convex function, $\mathbb{T} \subset \mathrm{dom}(\mathsf{F}_c)$ and for all $\gamma > 0$, the proximal map of $\gamma \mathsf{F}_c + \iota_{\mathbb{T}}$ is a point-to-point map with an explicit expression.

$\iota_{\mathbb{T}}$ denotes the characteristic function of the set $\mathbb{T}$; remember that the proximal map is given by

$$\mathsf{Prox}_{\gamma \mathsf{F}_c + \iota_{\mathbb{T}}}(\theta) := \mathrm{Argmin}_{\tau \in \mathbb{T}} \left(\gamma \mathsf{F}_c(\tau) + \frac{1}{2} \|\tau - \theta\|^2 \right).$$

The `Online Proximal-Gradient` algorithm, given by Algorithm 4, is a stochastic perturbation of a forward-backward splitting process in optimization (see Lions & Mercier (1979), see also (Beck & Teboulle, 2009, Section 1.3.3.) for an introduction as a `MM` technique) in which the forward operator is not explicit and learnt from a stream of oracles (see e.g. (Combettes & Wajs, 2005, Section 3) for a pioneering contribution on inexact proximal-gradient algorithms).

Algorithm 4 The `Online Proximal-Gradient` algorithm

Require: a sequence $\{N_t, t \geq 1\}$ of positive integers, an initial value $\theta^0 \in \mathbb{T}$, $\epsilon > 0$
Ensure: A $\mathbb{T}$-valued sequence $\{\theta^t, t \geq 0\}$.

1: **for** $t = 0, \cdots,$ **do**
2: Sample a minibatch $\{X^{t+1,i}, i = 1, \cdots, N_{t+1}\}$ of size N_{t+1}.
3: Set $\mathcal{H}^{t+1} := N_{t+1}^{-1} \sum_{i=1}^{N_{t+1}} \mathsf{h}\left(\theta^t, X^{t+1,i}\right)$
4: Choose $\gamma_{t+1} \in [\epsilon, 1/L]$
5: Compute

$$\theta^{t+1} = \mathsf{Prox}_{\gamma_{t+1} \mathsf{F}_c + \iota_{\mathbb{T}}}\left(\theta^t - \gamma_{t+1} \mathcal{H}^{t+1}\right). \tag{12}$$

6: **end for**

`Online Proximal-Gradient` **is a** `SAM2` **algorithm.** Set $\mathbb{U} := \mathbb{T} \times [\epsilon, 1/L]$ and for all $\theta \in \mathbb{T}$, set $\mathbb{U}[\theta] := \{(\theta, \gamma), \gamma \in [\epsilon, 1/L]\}$. The function

$$\mathsf{G}(\theta'; (\theta, \gamma)) := \mathsf{F}_s(\theta) + \langle \nabla \mathsf{F}_s(\theta), \theta' - \theta \rangle + \frac{1}{2\gamma} \|\theta' - \theta\|^2 + \mathsf{F}_c(\theta') + \iota_{\mathbb{T}}(\theta'), \qquad \theta' \in \mathbb{R}^d, \theta \in \mathbb{T}, \gamma \in [\epsilon, 1/L]$$

satisfies A1 under PG1. We also have $\mathsf{G}(\theta'; (\theta, \gamma)) = \mathbb{E}\left[\mathsf{g}(\theta', X; (\theta, \gamma))\right]$ where

$$\begin{aligned} \mathsf{g}(\theta', x; (\theta, \gamma)) &:= \mathsf{F}_s(\theta) + \langle \mathsf{h}(\theta, x), \theta' - \theta \rangle + \frac{1}{2\gamma} \|\theta' - \theta\|^2 + \mathsf{F}_c(\theta') + \iota_{\mathbb{T}}(\theta') \\ &= \mathsf{F}_c(\theta') + \frac{1}{2\gamma} \|\theta' - \theta + \gamma \mathsf{h}(\theta, x)\|^2 + \mathcal{C}(\gamma, \theta) + \iota_{\mathbb{T}}(\theta') \end{aligned}$$

where $\mathcal{C}(\gamma, \theta)$ does not depend on θ'. Therefore, the minimization (5) is equivalent to the computation of the proximal-gradient operator (12), thus showing that Algorithm 4 is a `SAM2` algorithm.

The limiting set. Under PG1 and

PG2. (a) F_s is convex,

(b) for all $x \in \mathbb{X}$, $\theta \mapsto \mathsf{h}(\theta, x)$ is continuous on $\mathbb{T}$ and $\mathbb{E}\left[\sup_{\theta \in \mathbb{T}} \|\mathsf{h}(\theta, X)\|^p\right] < \infty$ for some $p > 1$

the conditions A 1, A 2 and A 3 are satisfied. Since $\theta' \mapsto \mathsf{G}(\theta'; (\theta, \gamma))$ is a (proper) lower semicontinuous strictly convex function possessing a minimizer $\mathsf{T}(\theta, \gamma)$, then this minimizer is the unique solution of (see e.g. (Bauschke & Combettes, 2011, Theorem 16.3))

$$0 \in \partial \left[\langle \nabla \mathsf{F}_s(\theta), \cdot \rangle + \frac{1}{2\gamma} \| \cdot - \theta\|^2 + \mathsf{F}_c(\cdot) + \iota_{\mathbb{T}}(\cdot) \right] (\mathsf{T}(\theta, \gamma)).$$

By (Bauschke & Combettes, 2011, Corollary 16.48), $\mathsf{T}(\theta, \gamma) = \theta$, is equivalent to $0 \in \nabla \mathsf{F}_s(\theta) + \partial \mathsf{F}_c(\theta) + \partial \iota_{\mathbb{T}}(\theta)$. Observe that this property does not depend on γ which implies that $\mathcal{L}^+ = \mathcal{L}$. This property also yields

$$\mathcal{L} = \{\theta \in \mathbb{T} : 0 \in \partial \left[\mathsf{F} + \iota_{\mathbb{T}}\right](\theta)\}.$$

Since F is a lower semicontinuous proper convex function, then $\mathcal{L}$ is the set of the minimizers of F on $\mathbb{T}$.

For the sequence $\{\theta^t, t \geq 0\}$ given by Algorithm 4, Theorem 9 shows that as soon as the size of the mini batches is chosen so that A4 holds: the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ converges to $\min_{\mathbb{T}} \mathsf{F}$ and the sequence $\{\theta^t, t \geq 0\}$ converges to the set of the minimizers of F on $\mathbb{T}$.

In the literature, the convergence analysis of perturbed Proximal-Gradient algorithms usually relies on the Siegmund-Robbins lemma (see (Robbins & Siegmund, 1971, Theorem 1) for a stochastic version assuming the errors are nonnegative; see also Atchadé et al. (2017) for an extension addressing the case of signed errors and deterministic perturbations; and Lai (1989) for nonnegative errors and possibly non converging cumulated errors). Different settings were considered. Let

us cite as few examples, the finite-sum setting Nitanda (2014); Fort & Moulines (2023); the convex case under summability assumptions on the errors $\mathcal{H}^{t+1} - \nabla \mathsf{F}_s(\theta^t)$ which reveal restrictive for the i.i.d. streaming framework considered here Combettes & Pesquet (2015); the convex case under possibly biased errors Atchadé et al. (2017); the convex case when F_s is uniformly convex at a minimizer and the conditional variance of the error is controled by $\|\nabla \mathsf{F}_s(\theta^t)\|^2$ and does not vanish Rosasco et al. (2020). The closest result to the one we obtain is (Atchadé et al., 2017, Theorem 6), which covers Algorithm 4: applied with a lower bounded stepsize sequence $\{\gamma_t, t \geq 1\}$, it provides almost-sure convergence of the `SAM2` sequence to a point in $\mathcal{L}$ under the assumption $\sum_t N_t^{-1} < \infty$.

5 A numerical illustration: Regression under a quantile loss function

The effectiveness of the `SAM2` algorithm is demonstrated in the following simulation analysis. We consider a linear model with heavy-tailed noise, for which most existing theories and implementation are no longer applicable. In contrast, we show that `SAM2` applies straightforwardly with theoretical guarantees. More specifically, we consider the following linear model,

$$Y = \left\langle \theta, \overline{W} \right\rangle + \epsilon \tag{13}$$

where $\theta \in \mathbb{R}^\ell$ is the unknown regression parameter, $\overline{W} := (1, W)$ and $W \in \mathbb{R}^{\ell-1}$ are the covariates and $Y \in \mathbb{R}$ is the response variable. The random variables W and ϵ are defined in $(\Omega, \mathcal{A}, \mathbb{P})$; the additive noise ϵ is assumed independent of W and W is integrable $\mathbb{E}[\|W\|] < \infty$.

The goal is to learn the regression parameter θ from streaming data, independent and with the same distribution as $X := (W, Y)$. We are interested in very heavy-tailed noise ϵ, typically with infinite variance, for which the standard squared-loss approach cannot be applied: the quantile regression (QR) loss is commonly used instead. Let $q \in (0, 1)$ be a specified quantile level, the so-called QR problem consists of finding the solution of the following minimization problem

$$\arg\min_{\theta \in \mathbb{R}^\ell} \mathbb{E}\left[\rho_q(Y - \left\langle \theta, \overline{W} \right\rangle) - \rho_q(Y)\right], \qquad \rho_q(u) := (q - 1_{u<0})u. \tag{14}$$

This objective function is finite for all $\theta \in \mathbb{R}^\ell$ under the integrability condition on W (see e.g. Lemma 17). Note that the addition of the second term in (14) does not impact the solution of the minimization problem but guarantees the existence of the loss and allows to consider more general settings. In this section, we consider the possibly penalized QR problem and aim to solve:

$$\arg\min_{\theta} \mathsf{F}_\eta(\theta), \qquad \mathsf{F}_\eta(\theta) := \mathbb{E}\left[\rho_q(Y - \left\langle \theta, \overline{W} \right\rangle) - \rho_q(Y)\right] + \eta\|\theta\|_1, \tag{15}$$

where $\|\theta\|_1$ is the sparsity inducing L_1-norm of θ and $\eta \geq 0$. When $q = 1/2$, we recover the so-called

Least Absolute Deviation (LAD) problem. As the QR loss is non smooth, a number of standard algorithms cannot be applied. More specifically, the two main sources of complication in QR are the heavy-tail of the noise and the non-smoothness of the loss. The sometimes called *check* function ρ_q is non-differentiable at 0. When the noise is not of finite variance, most least-square-based theories or Huber-based robust approaches are not applicable. To handle non-smoothness, most approaches are two-step approaches with first the design of a smooth approximation of the QR loss and then the optimization of this approximation, with the hope that the distance between the estimates and the true minimizers can be controlled and tends to zero. Various smooth approximations have been proposed, using a perturbation of the check function that can be majorized by a quadratic funtion (Hunter & Lange, 2000), kernel convolution-based smoothing (Jiang & Yu, 2022; Chen et al., 2019), a huberized pinball loss for robust to outliers but essentially Gaussian noise (Ichinose et al., 2023), transformation of the QR loss into a least-square loss on new response variables (Chen et al., 2020), etc. The type of approximations designed generally guides in turn the subsequent optimization technique. For instance Hunter & Lange (2000) use an MM algorithm, while Chen et al. (2019, 2020); Ichinose et al. (2023); Zheng (2011) exploit the differentiability of their approximation to derive stochastic gradient or stochastic Newton-Raphson algorithms. The position of `SAM2` appears quite unique in the QR literature, as `SAM2` does not require to change the target loss to a smooth approximation while performing the loss optimization with an MM framework. The following proposition shows that we can indeed formulate the QR problem so as to exhibit a convenient majorizer, which is both easy to optimize and satisfies the `SAM2` requirements under mild conditions. Note however, that the result below only holds for linear regression functions while some of the previous mentioned methods, e.g. (Hunter & Lange, 2000; Ichinose et al., 2023) can also deal with non-linear regressions. Regarding the addition of a Lasso penalty for variable selection, most approaches including `SAM2` handle this extension straightforwardly.

Proposition 10. *Assume there exists $p_\star > 1$ such that $\mathbb{E}\left[\|W\|^{p_\star}\right] < \infty$. Then* A *1 and* A *2-Item a are satisfied with*

$$\mathsf{g}_\eta\left(\theta, X; \tau\right) := \frac{1}{\ell}\sum_{j=1}^{\ell}\Big(\rho_q(Y - \left\langle \tau, \overline{W}\right\rangle + \ell\,\overline{W}_j\tau_j - \ell\overline{W}_j\theta_j) - \rho_q(Y) + \eta\ell|\theta_j|\Big),$$

and $\mathbb{U}[\theta] = \{\theta\}$ for all θ. In addition, A *3 is satisfied with $p = p_\star$. The optimization step* (5) *is explicit for all $q \in (0,1)$ and $\eta \geq 0$.*

The proof is in Section 6.7. Whatever the dimension ℓ of θ, g_η is a sum of functions that involve each θ_i separately and can then be separately minimized over a one-dimensional space. A ℓ-dimensional problem is then turned into ℓ 1-dimensional problems. In Section 6.7, we show that such 1-dimensional minimization can be solved exactly.

5.1 Simulation Setup

Let $X^i := (W^i, Y^i) \in \mathbb{R}^{\ell-1} \times \mathbb{R}$ for each $i \in [N]$ be the realization of an i.i.d. sample of (W, Y). We consider the following linear model,

$$Y = \left\langle \theta^{\text{true}}, \overline{W} \right\rangle + \epsilon \tag{16}$$

where $\overline{W} := (1, W_1, \ldots W_{\ell-1})$ is a ℓ-dimensional covariate vector. The examples $\{W^i, i \in [N]\}$ are sampled independently, from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$ with a covariance matrix Σ whose entries are $\Sigma_{r,s} := r^{|r-s|}$ for $r, s \in [\ell - 1]$. We set $\ell = 11$ and $r = 0.9$. The true coefficient is set to

$$\theta^{\text{true}} := 10 \left(\frac{1}{\ell}, \frac{2}{\ell} \ldots, \frac{\ell - 1}{\ell}, 1 \right).$$

The noise distribution is chosen to be Cauchy $\epsilon \sim \text{Cauchy}(0, 1)$, and the random variables ϵ^i are independent.

The unknown parameter θ is learnt from the examples (X^i, Y^i), as a minimizer of the criterion (15) in the case $\eta = 0$ and $q = 0.5$. For `SAM2` the sequence $\{N_t, t \geq 1\}$ is chosen so as to satisfy assumption A4. For a Gaussian vector W, the assumption A3 is satisfied for any $p > 1$ so that the theory claims that a sufficient condition for the convergence of `SAM2` is the N_t's increase as slowly as desired. However, we illustrate below that too small constant values are not recommended. Thus, we first illustrate `SAM2` with a linearly growing modified so as to start with batch sizes of 100: $N_t = \max(100, t)$. As will be illustrated in Section 5.3, this burnin has no real impact on the algorithm convergence but starting with batch sizes of 100 provides better starts. The number of iterations is set to $T = 1000$. It follows that the total number of data points processed is $N = \sum_{t=1}^{T} N_t = 505,450$.

As the QR loss is non-smooth, for comparison, we implement a stochastic subgradient (`SSG`) algorithm using a subgradient $g_q(\theta; (w, y))$ of $\rho_q(y - \langle \theta, \bar{w} \rangle)$ with respect to θ which is given by

$$g_q(\theta; (w, y)) := -\left(q - 1_{y - \langle \theta, \bar{w} \rangle < 0}\right) \bar{w}; \qquad \bar{w} := (1, w).$$

The `SSG` algorithm is essentially the subgradient algorithm, but using noisy subgradients and a sequence of positive step sizes $\{\gamma_t, t \geq 1\}$. It uses the standard update specified in Algorithm 5, written with the possibility to use mini batches also denoted by $\{N_t, t \geq 1\}$.

`SSG` requires the choice of step sizes and mini-batch sizes, while `SAM2` requires only to choose a sequence of mini-batch sizes $\{N_t, t \geq 1\}$.

For all algorithms, the iterate is arbitrarily initialized to $\theta^0 := (1, \ldots, 1)$ and a Polyak averaging is also performed starting at $T_0 = 500$, for $t > T_0$: $\bar{\theta}^t = \frac{t-1-T_0}{t-T_0} \bar{\theta}^{t-1} + \frac{1}{t-T_0} \theta^t$. For a quantitative performance assessment, we compute the root mean squared error (RMSE) or L_2-error at each

Algorithm 5 The SSG algorithm

Require: a sequence $\{N_t, t \geq 1\}$ of positive integers and of positive step sizes $\{\gamma_t, t \geq 1\}$, an initial value $\theta^0 \in \mathbb{T}$

Ensure: A $\mathbb{T}$-valued sequence $\{\theta^t, t \geq 0\}$.

1: **for** $t = 0, \cdots,$ **do**

2: Sample a minibatch $\{X^{t+1,i}, i = 1, \cdots, N_{t+1}\}$ of size N_{t+1}.

3: Compute

$$\theta^{t+1} = \theta^t - \gamma_{t+1} \frac{1}{N_{t+1}} \sum_{i=1}^{N_{t+1}} g_q\left(\theta; X^{t+1,i}\right) \tag{17}$$

4: **end for**

iteration:

$$\text{RMSE}_t := ||\theta^t - \theta^{\text{true}}|| \, .$$

SAM2 is then compared to SSG using various settings in order to assess their respective robustness to the step size and batch size choices.

5.2 Sensitivity to step sizes

Regarding the SSG algorithm, we first consider a standard setting with mini-batches of equal size. To compare both algorithms using the same amount of observations and perform the same number of parameter updates, the batch size is set to $\lfloor N/T \rfloor = 505$ for all iterations except the last one which is augmented with the remaining data points and then includes 955 samples. This later adjustment has very little impact. SSG requires the user to decide on a step size sequence, which has been reported to be impactful and the main practical limitation of gradient algorithms. Three such sequences are thus tested: $\gamma_t = \frac{1}{(t+1)^{0.51}}$, $\gamma_t = \frac{1}{(t+1)^{0.6}}$ and $\gamma_t = \frac{1}{(t+1)^{0.7}}$.

Figure 1 shows, for one simulated data set, the sequences (left) of RMSE obtained with SAM2 and the three SSG tested settings, with a zoom on the last 500 iterations (right). Two representations of the RMSE evolution are reported in Figure 1, one with respect to the number of parameter updates (iterations, in the first line), and the other with respect to the number of processed samples (second line). The lowest RMSE are obtained for SAM2 and SSG with step sizes $\gamma_t = \frac{1}{(t+1)^{0.51}}$, with SAM2 outperforming the later. For $\gamma_t = \frac{1}{(t+1)^{0.7}}$, the SSG blue curve is not visible on the zoomed plot as its RMSE are too large. Polyak averaging performed with a burnin, starting at $T_0 = 500$, is beneficial to the SAM2 RMSE but not for the SSG RMSE. In terms of observed RMSE, the Polyak averaging performance decreases as the burnin phase T_0 becomes smaller. Note that the RSME curve of the Poliak averaging sequence is not the Polyak averaging of the RMSE sequence. Overall, although noisier, the SAM2 estimations appear less biased than the SSG ones. This is illustrated on some parameter sequences in Figure 2. The plots illustrate that SAM2 estimates are in general closer to the true values. In addition, the simulations confirm the often reported fact that SSG performance can be quite sensitive to the step size choice, with 2 out of 3 settings showing significantly larger

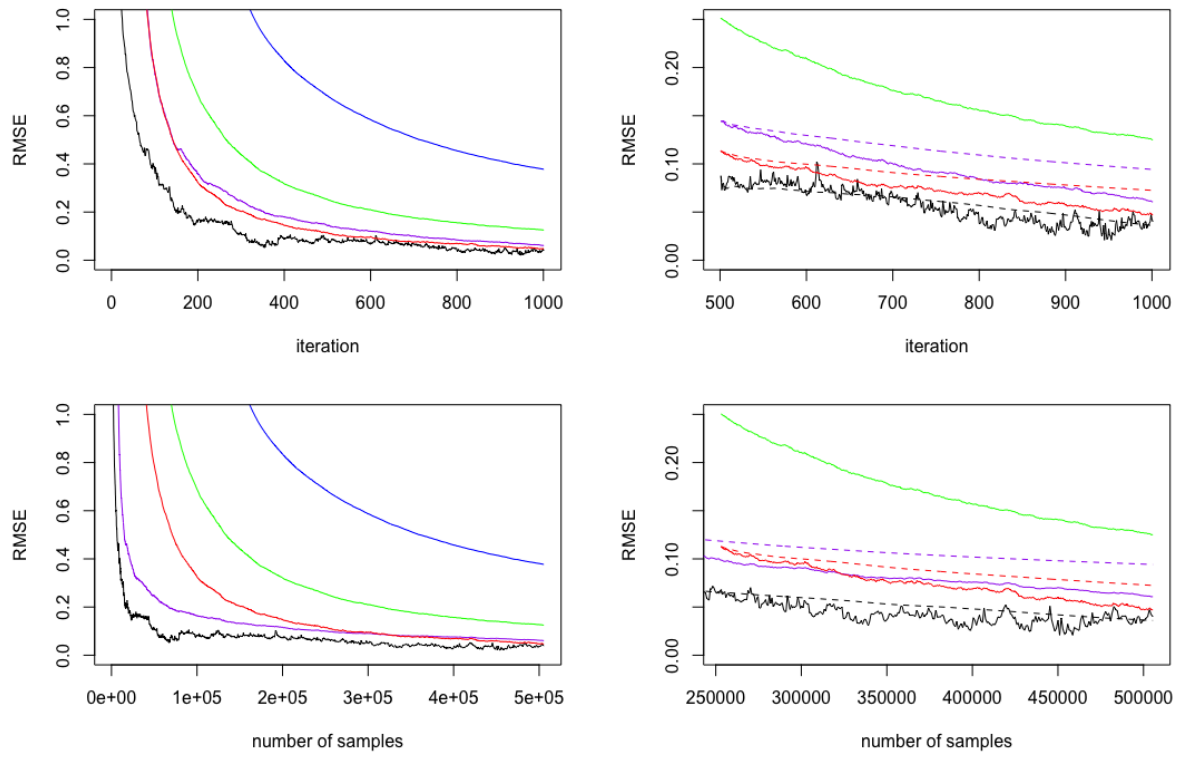

Figure 1: LAD example. First line: RMSE over iterations for SAM2 (black), SSG with different step size schedules, $\gamma_t = \frac{1}{(t+1)^{0.51}}$ (red and purple), $\gamma_t = \frac{1}{(t+1)^{0.6}}$ (green) and $\gamma_t = \frac{1}{(t+1)^{0.7}}$ (blue). All SSG sequences are with constant batch sizes except the purple curve which uses the same batches as SAM2. The left plot shows a zoom on the last half iterations, with additional RMSE curves for the Polyak averaging sequences (dashed lines). Second line: same comparison of RMSE but over number of processed samples.

RMSE. In contrast, SAM2 does not require such tuning. It is quite robust to the choice of the N_t's as soon as they are not kept too small for too many iterations. Typically, when $N_t = t$, which lowers the first batch sizes, SAM2 performs similarly as in Figure 1 but with a slower start. We further investigate the sensitivity to batch sizes in the next section.

5.3 Sensitivity to batch sizes

First, for SSG a second strategy is considered for the mini-batch size with a varying size N_t so that at each iteration the SSG algorithm uses the exact same data points as SAM2. This setting is tested only for SSG with $\gamma_t = \frac{1}{(t+1)^{0.51}}$, which provided before the best results. For this sequence of step sizes, the purple curve in Figure 1 shows RMSE obtained when using the same mini-batch sizes and data points for both SAM2 and SSG. The stronger impact is seen on the second line of Figure 1. SAM2 and SSG with linearly increasing batch sizes, show better start reaching lower RMSE faster. This is consistent with the intuition that more frequent parameter updating is preferable than finer optimization steps at the early stages of the algorithms. Overall, final SSG RMSE do not seem to be much impacted by this change of batch sizes. For this specific simulation, the constant batch

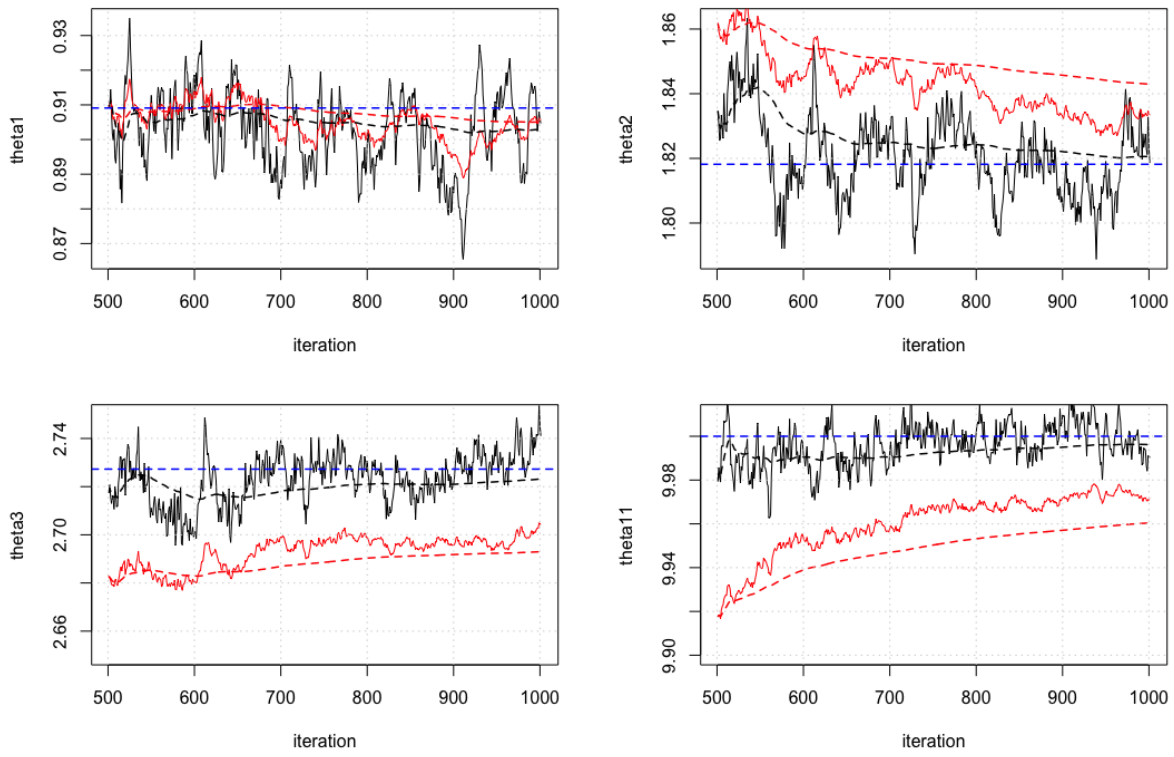

Figure 2: LAD example. Sequences of estimates for components $\theta_1, \theta_2, \theta_3$ and θ_{11}, using `SAM2` with linear $N_t = \max(100, t)$ (black) and `SSG` with $\gamma_t = \frac{1}{(t+1)^{0.51}}$ and constant batch sizes $N_t = 505$ (red). Polyak averaging sequences are shown with dashed lines. True parameter values are indicated by the blue horizontal dashed lines.

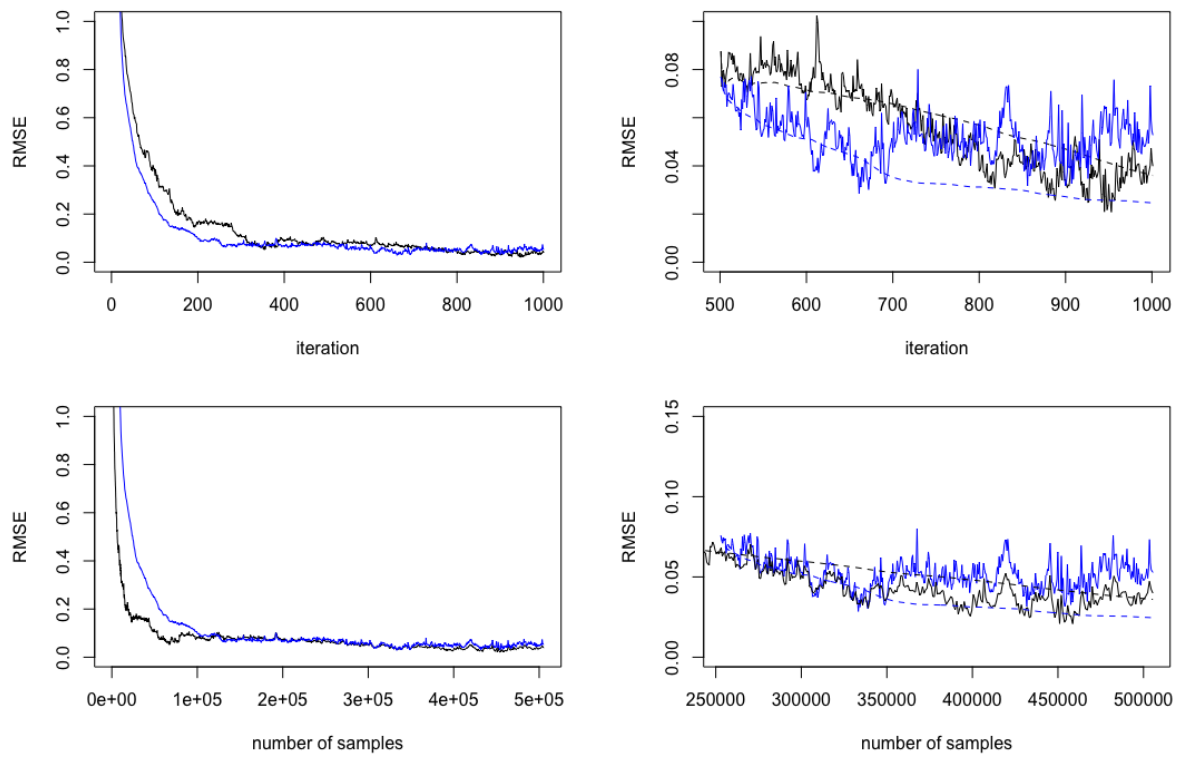

Figure 3: LAD example. First line: RMSE over iterations for SAM2 with different batch size sequences, original SAM2 (black) and constant (505) batch size SAM2 (blue). The left plot shows a zoom on the last half iterations, with additional RMSE curves for the Polyak averaging sequences (dashed lines). Second line: same comparison but with respect to the number of samples.

size setting slightly outperforms eventually the varying one but boxplots in Figure 5 show than on average both setting lead to very similar performance.

Symmetrically, Figure 3 shows then the result of two SAM2 runs, where the previous setting is compared with one where the batch sizes are set to a constant $\lfloor N/T \rfloor = 505$, as for the previous SSG. Although this does not satisfy condition A4 on the N_t's, this setting provides better RMSE in the first iterations but is eventually equivalent as the batch sizes increase. The Polyak averaging sequences suggest that the constant batch size version of SAM2 will eventually be outperformed by the increasing size one, as expected. Figure 4 further shows SAM2 sequences with constant batch sizes successively set to constant between 1 and 10. N_t should not be set too low. For $N_t = 1$, the algorithm diverges very quickly. For $N_t = 5$, RMSE values remain very high and noisy, while $N_t = 10$ still provides much larger RMSE than $N_t = 505$.

This first set of simulations suggests that SAM2 has the advantage, over its stochastic gradient SSG counterpart, not to require the choice of a step size sequence, while choosing batch sizes is not problematic. One could argue that the need to use increasing N_t's could be limiting, typically if one does not control the data stream. But note that this is only a sufficient condition and that SAM2 seems to perform well for constant N_t's too especially when combined with Polyak averaging. For SSG the sensitivity to the step size appears not to be as easily compensated by appropriate

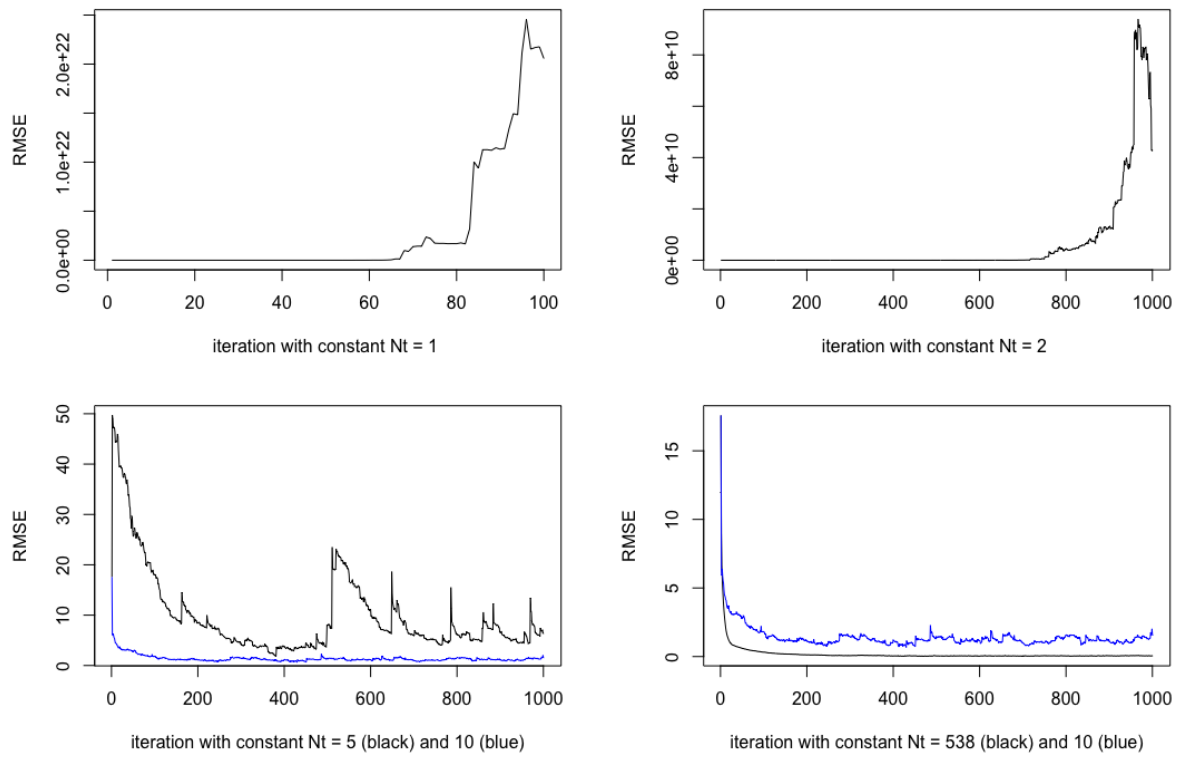

Figure 4: LAD example. RMSE over iterations for `SAM2` with constant batch size set from $N_t = 2$ to $N_t = 10$: (a) For $N_t = 1$ only the first 100 iterations are shown due to the explosion of the values, (b) For $N_t = 2$ note the very high RMSE values too, (c) Comparison of the $N_t = 5$ and $N_t = 10$, (d) $N_t = 10$ RMSE sequence and comparison with $N_t = 505$.

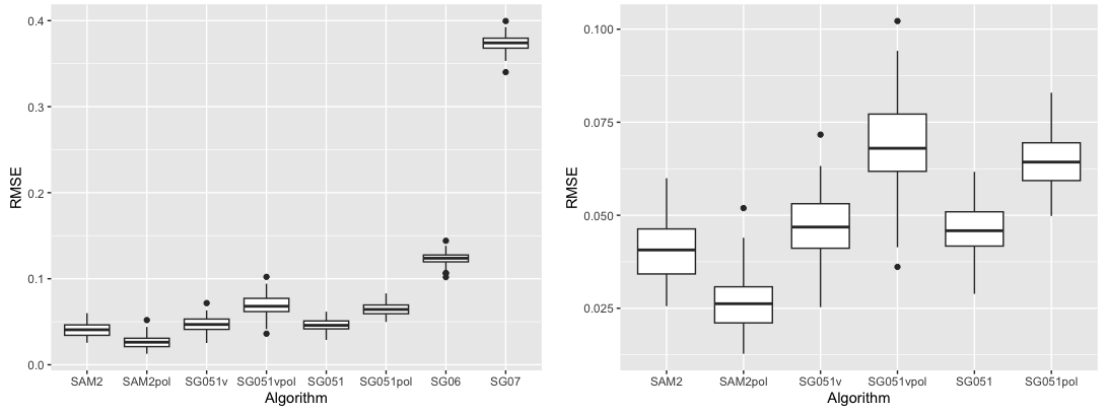

Figure 5: LAD example, 100 simulations. Last iteration RMSE boxplots for `SAM2` and `SSG` with different batch and step size schedules. Left plot: From left to right, `SAM2`, `SAM2` with Polyak averaging, `SSG` and its Poliak averaging, with varying batch size, with constant batch sizes set to 505, both with $\gamma_t = \frac{1}{(t+1)^{0.51}}$, `SSG` with constant batch sizes set to 505 with $\gamma_t = \frac{1}{(t+1)^{0.6}}$ and $\gamma_t = \frac{1}{(t+1)^{0.7}}$. Right plot: zoom on the 6 best algorithms. Poliak averaging is performed for all algorithms with $T_0 = 500$.

batch sizes.

5.4 Average assessment

The comparison is completed by repeating the same experiment for 100 different simulated data sets using the same QR model. Figure 5 shows the boxplots obtained with the `SAM2` and `SSG` algorithms, using the RMSE at the last iteration. More specifically, eight settings are considered: `SAM2` with $N_t = \max(100, t)$ and its Polyak averaging, `SSG` for $\gamma_t = \frac{1}{(t+1)^{0.51}}$ with both varying batch sizes and constant $N_t = 505$, and their respective Polyak averaging. `SSG` with $N_t = 505$ for $\gamma_t = \frac{1}{(t+1)^{0.6}}$ and $\gamma_t = \frac{1}{(t+1)^{0.7}}$ are also shown.

The same conclusions as in the previous sections hold. The sensitivity of `SSG` to step sizes is confirmed while `SAM2` outperforms the other settings with a simple increasing sequence of N_t's. `SAM2` produces noisier sequences which are easily smoothed via Polyak averaging, which further improves the final `SAM2` estimates. In contrast, Polyak averaging is not beneficial to `SSG` in none of the configurations. This is essentially due to the parameter sequences that are much more biased as illustrated in Figure 2. At last, the good performance of `SAM2` is at the expense of a higher computation cost due to the resolution of an optimization problem at each iteration, while `SSG` involves only a straightforward update.

6 Proofs and Technical results

6.1 Technical results on F, G and T

Proposition 11. *i) Assume* A *3-a.* G *is continuous on* $\mathbb{T}\times\mathbb{U}$*, and* $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ *is not empty for any* $\upsilon\in\mathbb{U}$*.*

ii) Assume A *3-c.* $\mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$ *is not empty.*

iii) Assume A *1. For any* $\upsilon\in\mathbb{U}$*,* $\min_{\mathbb{T}}\mathsf{F}\leq\min_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$*. In addition, for any* $\theta\in\mathbb{T}$ *and* $\upsilon\in\mathbb{U}[\theta]$*,* $0\leq\mathsf{G}(\theta;\upsilon)+\min_{\mathbb{T}}\mathsf{F}\leq\mathsf{F}(\theta)+\min_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$*.*

Proof. Under A3-a, the dominated convergence theorem implies that the function $(\theta,\upsilon)\to\mathsf{G}(\theta;\upsilon)$ is continuous. Since $\mathbb{T}$ is compact, $\inf_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ is not empty for any $\upsilon\in\mathbb{U}$.

Under A3-c, $\theta\mapsto\mathsf{F}(\theta)$ is continuous on the compact set $\mathbb{T}$; hence, $\mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$ is not empty.

Let $\upsilon\in\mathbb{U}$. A 1-a implies that $\min_{\mathbb{T}}\mathsf{F}\leq\min_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$. This inequality, combined with the equality in A1-b, yields the second statement. □

Proposition 12. *Assume* A *3-a and* A *3-b. Then*

i) T *is continuous on* $\mathbb{U}$*.*

ii) $(\theta,\upsilon)\mapsto\mathsf{G}(\theta;\upsilon)-\mathsf{G}(\mathsf{T}(\upsilon);\upsilon)$ *is continuous on* $\mathbb{T}\times\mathbb{U}$*.*

Proof. (i) Let $\upsilon^\star\in\mathbb{U}$ and $\{\upsilon^t,t\geq0\}$ be a $\mathbb{U}$-valued sequence such that $\lim_t\upsilon^t=\upsilon^\star$. Then, by A 3-b, for all $\theta\in\mathbb{T}$

$$\mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t)\leq\mathsf{G}(\theta;\upsilon^t).$$

By A 3-a, G is continuous on $\mathbb{T}\times\mathbb{U}$ (see Proposition 11); then $\lim_t\mathsf{G}(\theta;\upsilon^t)=\mathsf{G}(\theta;\upsilon^\star)$. The sequence $\{\mathsf{T}(\upsilon^t),t\geq0\}$ is a $\mathbb{T}$-valued sequence and there exists a subsequence $\{\upsilon^{\rho(t)},t\geq0\}$ such that $\lim_t\mathsf{T}(\upsilon^{\rho(t)})$ exists and is in $\mathbb{T}$. Note that, as a subsequence of $\{\upsilon^t,t\geq0\}$, we have $\lim_t\upsilon^{\rho(t)}=\upsilon^\star$. Using again that G is continuous, we have for all $\theta\in\mathbb{T}$

$$\mathsf{G}(\lim_t\mathsf{T}(\upsilon^{\rho(t)});\upsilon^\star)\leq\mathsf{G}(\theta;\upsilon^\star).$$

This inequality implies that $\lim_t\mathsf{T}(\upsilon^{\rho(t)})$ is a minimizer of $\theta\mapsto\mathsf{G}(\theta;\upsilon^\star)$. Since this minimizer is unique by A3-b, we have $\lim_t\mathsf{T}(\upsilon^{\rho(t)})=\mathsf{T}(\upsilon^\star)$. This holds true for any subsequence $\{\upsilon^{\rho(t)},t\geq0\}$ and concludes the proof.

(ii) It follows from the continuity of T and of G (see Proposition 11). □

6.2 Proof of Theorem 4

For all $\upsilon\in\mathbb{U}$, $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ is not empty by Proposition 11-i. $\mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$ is not empty by Proposition 11-ii.

Proof of Item ii. Let $\theta \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$. From (7), we have for any $\upsilon \in \mathbb{U}[\theta]$, $\min_{\mathbb{T}}\mathsf{F} = \mathsf{F}(\theta) \geq \mathsf{F}(\tau)$ for any $\tau \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$. This yields $\tau \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{F}$ and $\mathsf{F}(\theta) = \mathsf{F}(\tau)$. Hence $\theta \in \mathcal{L}$.
Let $\theta^\star \in \mathcal{L}$. Let $\upsilon \in \mathbb{U}[\theta^\star]$ and $\tau \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$. Since $\theta^\star \in \mathcal{L}$, we have $\mathsf{F}(\tau) = \mathsf{F}(\theta^\star)$. Combined with (7), this yields $\mathsf{G}(\theta^\star;\upsilon) = \mathsf{G}(\tau;\upsilon)$. Therefore, $\mathsf{G}(\theta^\star;\upsilon) = \min_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ and $\theta^\star \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$.
Proof of Item iii. By definition, we have $\mathcal{L} = \{\theta \in \mathbb{T} : \mathsf{F}(\mathsf{T}(\upsilon)) = \mathsf{F}(\theta) \text{ for all } \upsilon \in \mathbb{U}[\theta]\}$.
Let $\theta \in \mathbb{T}$ such that $\theta \in \mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon)$ for all $\upsilon \in \mathbb{U}[\theta]$. By A3-b, $\mathrm{Argmin}_{\mathbb{T}}\mathsf{G}(\cdot;\upsilon) = \{\mathsf{T}(\upsilon)\}$ which implies that $\theta = \mathsf{T}(\upsilon)$ for all $\upsilon \in \mathbb{U}[\theta]$. Hence, $\mathsf{F}(\theta) = \mathsf{F}(\mathsf{T}(\upsilon))$ for all $\upsilon \in \mathbb{U}[\theta]$ and $\theta \in \mathcal{L}$. The proof is concluded with (7).

6.3 Proof of Theorem 5

Since F is continuous and $\mathcal{L} \cap \mathbb{T}$ is compact, the set $\mathcal{D} := \mathsf{F}(\mathcal{L} \cap \mathbb{T})$ is a compact subset of $\mathbb{R}$. Denote by $\mathcal{D}_\beta$ the β-neighborhood of the closed set $\mathcal{D}$: $\mathcal{D}_\beta := \{w \in \mathbb{R} : \mathrm{d}(w, \mathcal{D}) < \beta\}$; d is the Euclidean distance from a point to a closed set. Since $\mathcal{D}$ is compact, we have

$$\mathcal{D} = \bigcap_{\beta>0} \mathcal{D}_\beta. \tag{18}$$

Step 1. Let $\alpha > 0$. Since $\mathcal{D}_\alpha$ is a finite union of disjoint bounded intervals, there exist a positive integer n_α and two increasing real-valued sequences $\{a_\alpha(k), 1 \leq k \leq n_\alpha\}$ and $\{b_\alpha(k), 1 \leq k \leq n_\alpha\}$ such that

$$\mathcal{D}_\alpha = \bigcup_{k=1}^{n_\alpha} \left(a_\alpha(k), b_\alpha(k)\right). \tag{19}$$

Step 2. Since F is continuous and $\mathcal{D}_{\alpha/2}$ is open, $\mathsf{F}^{-1}(\mathcal{D}_{\alpha/2})$ is an open covering of $\mathcal{L} \cap \mathbb{T}$. Define

$$\epsilon_\alpha := \inf_{\theta \in \mathbb{T} \setminus \mathsf{F}^{-1}(\mathcal{D}_{\alpha/2}), \upsilon \in \mathbb{U}[\theta]} \left(\mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\upsilon))\right); \qquad \rho_\alpha := \epsilon_\alpha \wedge \alpha. \tag{20}$$

Observe that since $\mathbb{T} \setminus \mathsf{F}^{-1}(\mathcal{D}_{\alpha/2})$ is a compact subset of $\mathbb{T} \setminus \mathcal{L}$, then $\epsilon_\alpha > 0$ (and therefore $\rho_\alpha > 0$) by H-ii.

Step 3. We write

$$\mathsf{F}(\theta^t) - \mathsf{F}(\theta^{t+1}) = \mathsf{F}(\theta^t) - \mathsf{F}(\mathsf{T}(\upsilon^t)) + \mathsf{F}(\mathsf{T}(\upsilon^t)) - \mathsf{F}(\theta^{t+1}). \tag{21}$$

By H-iii, there exists T_α such that

$$\forall t \geq T_\alpha, \qquad |\mathsf{F}(\mathsf{T}(\upsilon^t)) - \mathsf{F}(\theta^{t+1})| \leq \rho_\alpha/2. \tag{22}$$

Hence, if $t \geq T_\alpha$ and $\theta^t \in \mathbb{T} \setminus \mathsf{F}^{-1}(\mathcal{D}_{\alpha/2})$ then (20)-(22) imply that $\mathsf{F}(\theta^t) - \mathsf{F}(\theta^{t+1}) \geq \epsilon_\alpha - \rho_\alpha/2 \geq \rho_\alpha/2$. Consequently, upon noting that $\inf_{\mathbb{T}} \mathsf{F} > -\infty$, the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ is infinitely often in $\mathcal{D}_{\alpha/2}$ and therefore in $\mathcal{D}_\alpha$. This implies that the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ is infinitely often in an interval of

(19): set $I_\alpha := (a_\alpha(k_\star), b_\alpha(k_\star))$ where $k_\star$ is defined by $k_\star := \max\{k \in \{1, \cdots, n_\alpha\} : \liminf_t \mathsf{F}(\theta^t) > a_\alpha(k)\}$.

Step 4. Let $t \geq T_\alpha$ such that $\mathsf{F}(\theta^t) \in I_\alpha$. We prove by induction that for any $s \geq 0$, $\mathsf{F}(\theta^{t+s}) < b_\alpha(k_\star)$. The property holds true for $s = 0$; assume it holds for $s > 0$. Using again (21), we distinguish two cases

- if $\mathsf{F}(\theta^{t+s}) \in \mathcal{D}_\alpha \setminus \mathcal{D}_{\alpha/2}$, then (20)-(22) imply that $\mathsf{F}(\theta^{t+s}) - \mathsf{F}(\theta^{t+s+1}) \geq \epsilon_\alpha - \rho_\alpha/2 \geq \rho_\alpha/2$. Hence $\mathsf{F}(\theta^{t+s+1}) \leq \mathsf{F}(\theta^{t+s}) - \rho_\alpha/2 < b_\alpha(k_\star) - \rho_\alpha/2 < b_\alpha(k_\star)$.
- if $\mathsf{F}(\theta^{t+s}) \in \mathcal{D}_{\alpha/2}$, then (21)-(22) and H-i imply that $\mathsf{F}(\theta^{t+s}) - \mathsf{F}(\theta^{t+s+1}) \geq 0 - \rho_\alpha/2$. Since $\mathsf{F}(\theta^{t+s}) < b_\alpha(k_\star) - \alpha/2$ and $\rho_\alpha \leq \alpha$, then $\mathsf{F}(\theta^{t+s+1}) < b_\alpha(k_\star)$.

This concludes the induction.

Step 5. We have $\liminf_t \mathsf{F}(\theta^t) > a_\alpha(k_\star)$ and there exists t such that for all $\tau \geq 0$, $\mathsf{F}(\theta^{t+\tau}) < b_\alpha(k_\star)$. Hence, $\mathsf{F}(\theta^t) \in I_\alpha$ for any t large enough: the set $\mathcal{F}$ of the limiting points of $\{\mathsf{F}(\theta^t), t \geq 0\}$ is not empty and is included in the interval I_α.

Let $0 < \alpha_1 < \alpha_2$: $I_{\alpha_1} \subseteq I_{\alpha_2}$ so that for any positive sequence $\{\alpha_n, n \geq 0\}$ decreasing to zero, $\mathcal{F} \subset \bigcap_n I_{\alpha_n}$ and $\bigcap_n I_{\alpha_n}$ is an interval. Note also that $\bigcap_n I_{\alpha_n} \subset \mathsf{F}(\mathcal{L} \cap \mathbb{T})$ by (18). Therefore, the sequence $\{\mathsf{F}(\theta^t), t \geq 0\}$ converges to an interval in $\mathsf{F}(\mathcal{L} \cap \mathbb{T})$; this concludes the proof of the first claim.

Step 6. Let us prove the last statement. If $\mathsf{F}(\mathcal{L} \cap \mathbb{T})$ has an empty interior, there exists $\mathsf{F}^\star$ such that $\lim_t \mathsf{F}(\theta^t) = \mathsf{F}^\star$. Since $\{\theta^t, t \geq 0\}$ is a compact sequence and F is continuous, then $\{\theta^t, t \geq 0\}$ converges to the set $\{\theta \in \mathbb{T} : \mathsf{F}(\theta) = \mathsf{F}^\star\}$.

Let a converging subsequence $\{\theta^{\rho(t)}, t \geq 0\}$, with limiting value $\theta^\star \in \mathbb{T}$; such a sequence exists since $\mathbb{T}$ is compact. The proof is by contradiction: assume that $\theta^\star \notin \mathcal{L}$. Since $\mathcal{L} \cap \mathbb{T}$ is closed, there exists $\delta > 0$ such that the compact ball $\mathcal{K}_\delta := \{\theta \in \mathbb{T} : \|\theta - \theta^\star\| \leq \delta\}$ is in $\mathbb{T} \setminus \mathcal{L}$. The assumption H-ii implies that there exists $\epsilon_\delta > 0$ such that for all $\theta \in \mathcal{K}_\delta$ and for **all** $\upsilon \in \mathbb{U}[\theta]$, $\mathsf{F}(\theta) > \epsilon_\delta + \mathsf{F}(\mathsf{T}(\upsilon))$. Since $\lim_t \theta^{\rho(t)} = \theta^\star$, there exists T_δ such for all $t \geq T_\delta$, $\theta^{\rho(t)} \in \mathcal{K}_\delta$. Therefore for all $t \geq T_\delta$, $\mathsf{F}(\theta^{\rho(t)}) > \epsilon_\delta + \mathsf{F}(\mathsf{T}(\upsilon^{\rho(t)}))$ which yields

$$\lim_t \mathsf{F}(\theta^{\rho(t)}) \geq \epsilon_\delta + \lim_t \mathsf{F}(\mathsf{T}(\upsilon^{\rho(t)})). \tag{23}$$

On the other hand, the assumption H-iii and the result $\lim_t \mathsf{F}(\theta^t) = \mathsf{F}^\star$ imply that

$$\lim_t \mathsf{F}(\mathsf{T}(\upsilon^{\rho(t)})) = \lim_t \mathsf{F}(\theta^{\rho(t)+1}), \qquad \mathsf{F}^\star = \lim_t \mathsf{F}(\theta^t) = \lim_t \mathsf{F}(\theta^{\rho(t)}) = \lim_t \mathsf{F}(\theta^{\rho(t)+1}). \tag{24}$$

The results (23) and (24) imply that $\mathsf{F}^\star \geq \epsilon_\delta + \mathsf{F}^\star$ which is a contradiction.

6.4 Proof of Proposition 7

Let $\epsilon > 0$. Since F is continuous (see A3-c) and $\mathbb{T}$ is compact: there exists $\eta_\epsilon > 0$ such that

$$\|\tau - \theta'\| \leq \eta_\epsilon \Longrightarrow |\mathsf{F}(\tau) - \mathsf{F}(\theta')| \leq \epsilon/2.$$

and $\sup_{\mathbb{T}} |\mathsf{F}| < \infty$. It holds

$$\begin{aligned} |\mathsf{F}(\mathsf{T}(\upsilon)) - \mathsf{F}(\theta')| &= |\mathsf{F}(\mathsf{T}(\upsilon)) - \mathsf{F}(\theta')| \left(1_{\|\mathsf{T}(\upsilon)-\theta'\| \leq \eta_\epsilon} + 1_{\|\mathsf{T}(\upsilon)-\theta'\| > \eta_\epsilon}\right) \\ &\leq \frac{\epsilon}{2} + \frac{2}{\eta_\epsilon} \sup_{\mathbb{T}} |\mathsf{F}| \ \|\mathsf{T}(\upsilon) - \theta'\|. \end{aligned}$$

Set $\delta_\epsilon := \epsilon\eta_\epsilon/(4 \sup_{\mathbb{T}} |\mathsf{F}|)$. Then we write by using (9)

$$\begin{aligned} \|\mathsf{T}(\upsilon) - \theta'\| &\leq \|\mathsf{T}(\upsilon) - \theta'\| 1_{\|\mathsf{T}(\upsilon)-\theta'\| < \delta_\epsilon} + \|\mathsf{T}(\upsilon) - \theta'\| 1_{\|\mathsf{T}(\upsilon)-\theta'\| \geq \delta_\epsilon} \\ &\leq \delta_\epsilon + \|\mathsf{T}(\upsilon) - \theta'\| 1_{\|\mathsf{T}(\upsilon)-\theta'\| \geq \delta_\epsilon} \\ &\leq \delta_\epsilon + 2 \ \mathrm{Diam}(\mathbb{T}) \ 1_{\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon) \geq \tilde{\eta}_{\delta_\epsilon}} \\ &\leq \delta_\epsilon + \frac{2}{\tilde{\eta}_{\delta_\epsilon}} \ \mathrm{Diam}(\mathbb{T}) \ \Big(\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon)\Big). \end{aligned}$$

As a conclusion, by using the definition of δ_ϵ, we have

$$|\mathsf{F}(\mathsf{T}(\upsilon)) - \mathsf{F}(\theta')| \leq \epsilon + \frac{4}{\eta_\epsilon \ \tilde{\eta}_{\delta_\epsilon}} \ \sup_{\mathbb{T}} |\mathsf{F}| \ \mathrm{Diam}(\mathbb{T}) \ \Big(\mathsf{G}(\theta';\upsilon) - \mathsf{G}(\mathsf{T}(\upsilon);\upsilon)\Big).$$

6.5 Uniform strong Law of Large Numbers for double arrays

Theorem 13 establishes a uniform strong Law of Large Numbers for double arrays, with the proof following the same lines of argument as that of (Andrews, 1992, Thm. 3), for example.

Theorem 13. *Assume*

B1. i) $\mathbb{V}$ is a compact subset of $\mathbb{R}^\upsilon$ and $(\mathbb{X}, \mathcal{X})$ is a measure space.

ii) $\mathsf{h} : \mathbb{V} \times \mathbb{X} \to \mathbb{R}$ is measurable and there exists $p > 1$ such that $\mathbb{E}\left[\sup_{u \in \mathbb{V}} |\mathsf{h}(u, X)|^p\right] < \infty$, where X is a $\mathbb{X}$-valued random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$.

iii) For all $x \in \mathbb{X}$, $u \mapsto \mathsf{h}(u, x)$ is continuous on $\mathbb{V}$.

iv) The deterministic $\mathbb{Z}_{>0}$-valued sequence $\{N_t, t \geq 1\}$ satisfies $\sum_{t \geq 1} N_t^{-((p-1)\wedge(p/2))} < \infty$.

v) The random variables $\{X^{t,i}, t \geq 1, 1 \leq i \leq N_t\}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$, are i.i.d. with the same distribution as X.

Then, with probability one,

$$\lim_{t\to\infty}\sup_{u\in\mathbb{V}}\left|\frac{1}{N_t}\sum_{i=1}^{N_t}\mathsf{h}\left(u,X^{t,i}\right)-\mathbb{E}\left[\mathsf{h}\left(u,X\right)\right]\right|=0.$$

Set $\mathsf{H}(u) := \mathbb{E}\left[\mathsf{h}(u,X)\right]$. Observe that under B 1-ii and B 1-iii, the dominated convergence theorem implies that H is continuous on $\mathbb{V}$.

The proof of Theorem 13 is based on Proposition 14: statement i) establishes a strong Law of Large numbers for a fixed $u \in \mathbb{V}$, and statement ii) establishes a strong Law of Large numbers for the modulus of continuity defined for any $\delta > 0$, by

$$\omega_\delta(x) := \sup_{u\in\mathbb{V},u'\in\mathbb{V},\|u-u'\|\leq\delta}|\mathsf{h}(u,x)-\mathsf{h}(u',x)|.$$

In the terminology of Davidson (2021, Sec. 22.4), statement i) is referred to as almost sure pointwise convergence of the quantity $N_t^{-1}\sum_{i=1}^{N_t}\mathsf{h}\left(u,X^{t,i}\right)-\mathsf{H}\left(u\right)$ to zero, and statement ii) implies its strong asymptotic uniform stochastic equicontinuity (often shorted to strong stochastic equicontinuity, for brevity), via the Markov's inequality. By Davidson (2021, Thm. 22.8), when $\mathbb{V}$ is compact (and thus totally bounded and separable), taken together, almost sure pointwise convergence and strong stochastic equicontinuity are equivalent to almost sure uniform convergence to zero of $N_t^{-1}\sum_{i=1}^{N_t}\mathsf{h}\left(\cdot,X^{t,i}\right)-\mathsf{H}$. Thus, under A6, the conclusions of Theorem 13 and Proposition 14 are almost equivalent.

Proposition 14. *Assume* B *1.*

i) For any $u \in \mathbb{V}$, there exists $\Omega_u \in \mathcal{A}$ such that $\mathbb{P}(\Omega_u) = 1$ and on Ω_u

$$\lim_{t\to\infty}\left|\frac{1}{N_t}\sum_{i=1}^{N_t}\mathsf{h}\left(u,X^{t,i}\right)-\mathsf{H}(u)\right|=0.$$

ii) For any $\delta > 0$, $\mathbb{E}[(\omega_\delta(X))^p] < \infty$ and there exists $\bar{\Omega}_\delta \in \mathcal{A}$ such that $\mathbb{P}(\bar{\Omega}_\delta) = 1$ and on $\bar{\Omega}_\delta$

$$\lim_{t\to\infty}\left|\frac{1}{N_t}\sum_{i=1}^{N_t}\omega_\delta(X^{t,i})-\mathbb{E}\left[\omega_\delta(X)\right]\right|=0.$$

Finally, $\lim_{\delta\to0}\mathbb{E}\left[\omega_\delta(X)\right]=0$.

Proof. Item i. Let $u \in \mathbb{V}$. Set $Y^{t,i} := \mathsf{h}\left(u,X^{t,i}\right)-\mathsf{H}(u)$. We prove that for any $\varepsilon > 0$, $\mathbb{P}(\limsup_t A_\varepsilon^t) = 0$ where $A_\varepsilon^t := \{\ N_t^{-1}\left|\sum_{i=1}^{N_t}Y^{t,i}\right| \geq \varepsilon\}$. This will imply that with probability one, $\lim_{t\to\infty}N_t^{-1}\sum_{i=1}^{N_t}Y^{t,i}=0$ and conclude the proof.

We write, by using the Markov inequality,

$$\sum_{t\geq 1}\mathbb{P}(A_\varepsilon^t)=\sum_{t\geq 1}\mathbb{P}\left(\left|N_t^{-1}\sum_{i=1}^{N_t}Y^{t,i}\right|\geq\varepsilon\right)\leq\varepsilon^{-p}\sum_{t\geq 1}N_t^{-p}\mathbb{E}\left[\left|\sum_{i=1}^{N_t}Y^{t,i}\right|^p\right]$$

where $p>1$ is given by B1-ii. The random variables $\{Y^{t,i},t\geq 1,1\leq i\leq N_t\}$ are i.i.d. and centered under B1-ii and B1-v. Hence

$$\mathbb{E}\left[\left|\sum_{i=1}^{N_t}Y^{t,i}\right|^p\right]=\mathbb{E}\left[\left|\sum_{i=1}^{N_t}Y^{\star,i}\right|^p\right],$$

where $\{Y^{\star,i},i\geq 1\}$ are i.i.d. random variables with the same distribution as $Y^{1,1}$. $n\mapsto\sum_{i=1}^n Y^{\star,i}$ is a martingale sequence with L^p-moment (see B1-ii): by (Hall & Heyde, 1980, Section 2.4 Theorem 2.10), there exists a constant C_p such that for any $t\geq 1$,

$$\mathbb{E}\left[\left|\sum_{i=1}^{N_t}Y^{\star,i}\right|^p\right]\leq C_p\,\mathbb{E}\left[\left|\sum_{i=1}^{N_t}|Y^{\star,i}|^2\right|^{p/2}\right].$$

From standard calculations (see e.g. Lemma 15), this yields

$$\mathbb{E}\left[\left|\sum_{i=1}^{N_t}Y^{\star,i}\right|^p\right]\leq C_p\,N_t^{(p/2)\vee 1}\,\mathbb{E}\left[|Y^{\star,1}|^p\right].$$

Under B1-iv, the Borel-Cantelli lemma implies that $\mathbb{P}(\limsup_t A_\varepsilon^t)=0$.

Item ii. Let $\delta>0$. We write $\omega_\delta(x)\leq 2\sup_{u\in\mathbb{V}}|\mathsf{h}(u,x)|$; from B1-ii, we have $\mathbb{E}\left[(\omega_\delta(X))^p\right]<\infty$.

Set $Y^{t,i}:=\omega_\delta(X^{t,i})-\mathbb{E}\left[\omega_\delta(X)\right]$. Following the same lines as in the proof of Item i, it can be proved that under B1-ii, B1-iv and B1-v, $\lim_t N_t^{-1}\sum_{i=1}^{N_t}Y^{t,i}=0$ with probability one; details are omitted.

Let us apply the dominated convergence theorem. Let $x\in\mathbb{X}$. By B1-i and B1-iii, for any $\epsilon>0$, there exists $\eta_{\epsilon,x}>0$ such that $\|u-u'\|\leq\eta_{\epsilon,x}$ implies $|\mathsf{h}(u,x)-\mathsf{h}(u',x)|\leq\epsilon$. Therefore, for any $\epsilon>0$, there exists $\eta_{\epsilon,x}>0$ such that for any $\delta\in[0,\eta_{\epsilon,x}]$, $\omega_\delta(x)\leq\epsilon$. This implies that $\lim_{\delta\to 0}\omega_\delta(x)=0$. In addition, $\omega_\delta(X)\leq 2\sup_{u\in\mathbb{V}}|\mathsf{h}(u,X)|$ and the RHS is integrable by B1-ii. Therefore, the assumptions of the dominated convergence Theorem are satisfied and $\lim_{\delta\to 0}\mathbb{E}\left[\omega_\delta(X)\right]=0$. □

Proof of Theorem 13. Define $\mathbb{I}_\mathbb{Q}:=\{1/q,q\in\mathbb{Z}_{>0}\}$. Since $\mathbb{V}$ is compact (see B1-i), for any $\delta\in\mathbb{I}_\mathbb{Q}$, there exists a finite covering of $\mathbb{V}$ by closed balls $\{B_{\delta,1},\cdots,B_{\delta,L_\delta}\}$ of radius δ; let us denote by $\upsilon_{\delta,1},\cdots,\upsilon_{\delta,L_\delta}$ the centers of these balls. Set

$$\Omega^\star:=\Big(\bigcap_{\delta\in\mathbb{I}_\mathbb{Q},\ell\in\{1,\cdots,L_\delta\}}\Omega_{\upsilon_{\delta,\ell}}\Big)\cap\Big(\bigcap_{\delta\in\mathbb{I}_\mathbb{Q}}\bar{\Omega}_\delta\Big),$$

where Ω_υ and $\bar{\Omega}_\delta$ are given respectively by Proposition 14-i and Proposition 14-ii. Since $\Omega^\star$ is a countable intersection of sets of probability one, then $\mathbb{P}(\Omega^\star) = 1$. We prove that on $\Omega^\star$,

$$\lim_{t\to\infty} \sup_{u\in\mathbb{V}} \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u, X^{t,i}\right) - \mathsf{H}(u) \right| = 0.$$

Let $\epsilon > 0$. By Item ii, there exists $\delta_\epsilon > 0$ such that $|\mathbb{E}\left[\omega_{\delta_\epsilon}(X)\right]| \leq \epsilon/6$; without loss of generality we can assume that $\delta_\epsilon \in \mathbb{I}_\mathbb{Q}$ and we do so. We write

$$\sup_{u\in\mathbb{V}} \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u, X^{t,i}\right) - \mathsf{H}(u) \right| \leq \sup_{\ell\in\{1,\ldots,L_{\delta_\epsilon}\}} \sup_{u\in B_{\delta_\epsilon,\ell}} \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u, X^{t,i}\right) - \mathsf{H}(u) \right|.$$

We now prove that on $\Omega^\star$, there exists T_ϵ such that for all $t \geq T_\epsilon$, each of the L_{δ_ϵ} terms in the RHS is upper bounded by ϵ. This will conclude the proof.

Fix $\ell \in \{1, \cdots, L_{\delta_\epsilon}\}$. We write $\sup_{u\in B_{\delta_\epsilon,\ell}} \left| N_t^{-1} \sum_{i=1}^{N_t} \mathsf{h}\left(u, X^{t,i}\right) - \mathsf{H}(u) \right| \leq \sum_{j=1}^{2} \mathcal{T}_{j,\epsilon,\ell}(t) + \mathcal{T}_{3,\epsilon,\ell}$ where

$$\begin{aligned}
\mathcal{T}_{1,\epsilon,\ell}(t) &:= \sup_{u\in B_{\delta_\epsilon,\ell}} \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u, X^{t,i}\right) - \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u_{\delta_\epsilon,\ell}, X^{t,i}\right) \right|, \\
\mathcal{T}_{2,\epsilon,\ell}(t) &:= \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u_{\delta_\epsilon,\ell}, X^{t,i}\right) - \mathsf{H}(u_{\delta_\epsilon,\ell}) \right|, \\
\mathcal{T}_{3,\epsilon,\ell} &:= \sup_{u\in B_{\delta_\epsilon,\ell}} \left| \mathsf{H}(u_{\delta_\epsilon,\ell}) - \mathsf{H}(u) \right|.
\end{aligned}$$

It holds by definitions of ω_δ and δ_ϵ,

$$\mathcal{T}_{1,\epsilon,\ell}(t) \leq \frac{1}{N_t} \sum_{i=1}^{N_t} \omega_{\delta_\epsilon}(X^{t,i}) \leq \frac{\epsilon}{6} + \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \omega_{\delta_\epsilon}(X^{t,i}) - \mathbb{E}\left[\omega_{\delta_\epsilon}(X)\right] \right|;$$

by Item ii, since $\delta_\epsilon \in \mathbb{I}_\mathbb{Q}$, on $\Omega^\star$ there exists $T_{1,\epsilon}$ such that for any $t \geq T_{1,\epsilon}$, we have $\mathcal{T}_{1,\epsilon,\ell} \leq \epsilon/3$.

By Item i, on $\Omega^\star$ there exists $T_{2,\epsilon}$ such that for any $t \geq T_{2,\epsilon}$, $\mathcal{T}_{2,\epsilon,\ell}(t) \leq \epsilon/3$.

Finally, $\mathcal{T}_{3,\epsilon,\ell} \leq \mathbb{E}\left[\omega_{\delta_\epsilon}(X)\right] \leq \epsilon/6$ by definition of δ_ϵ.

Set $T_\epsilon := T_{1,\epsilon} \vee T_{2,\epsilon}$. On $\Omega^\star$, for any $t \geq T_\epsilon$ and for any $\ell \in \{1, \cdots, L_{\delta_\epsilon}\}$,

$$\sup_{u\in B_{\delta_\epsilon,\ell}} \left| \frac{1}{N_t} \sum_{i=1}^{N_t} \mathsf{h}\left(u, X^{t,i}\right) - \mathsf{H}(u) \right| \leq \frac{2\epsilon}{3} + \frac{\epsilon}{6} \leq \epsilon.$$

Lemma 15. *Let $\{Y^i, i \geq 1\}$ be i.i.d. variables such that $\mathbb{E}\left[|Y^1|^p\right] < \infty$ for some $p > 1$. Then,*

$$\mathbb{E}\left[\left|\sum_{i=1}^{N}(Y^i)^2\right|^{p/2}\right] \leq N^{(p/2)\vee 1}\ \mathbb{E}\left[|Y^1|^p\right].$$

Proof. First case: $p/2 \leq 1$. We write

$$\mathbb{E}\left[\left|\sum_{i=1}^{N}(Y^i)^2\right|^{p/2}\right] \leq \mathbb{E}\left[\sum_{i=1}^{N}\left|(Y^i)^2\right|^{p/2}\right] = \mathbb{E}\left[\sum_{i=1}^{N}\left|Y^i\right|^{p}\right] = N\ \mathbb{E}\left[|Y^1|^p\right].$$

Second case: $p/2 \geq 1$. By the Minkowski inequality, we write

$$\mathbb{E}\left[\left|\sum_{i=1}^{N}(Y^i)^2\right|^{p/2}\right] = \left(\left\|\sum_{i=1}^{N}(Y^i)^2\right\|_{p/2}\right)^{p/2} \leq \left(\sum_{i=1}^{N}\left\|(Y^i)^2\right\|_{p/2}\right)^{p/2}.$$

The RHS is equal to

$$\left(\sum_{i=1}^{N}\mathbb{E}\left[|Y^i|^p\right]^{2/p}\right)^{p/2} = \left(\sum_{i=1}^{N}\mathbb{E}\left[|Y^1|^p\right]^{2/p}\right)^{p/2} = \left(N\ \mathbb{E}\left[|Y^1|^p\right]^{2/p}\right)^{p/2} = N^{p/2}\,\mathbb{E}\left[|Y^1|^p\right].$$

This concludes the proof. □

6.6 Proof of Theorem 9

6.6.1 Proof of Item i

We write

$$\begin{aligned}
&\mathsf{G}(\theta^{t+1};\upsilon^t) - \mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t) \\
&\qquad \leq \mathsf{G}(\theta^{t+1};\upsilon^t) - \frac{1}{N_{t+1}}\sum_{i=1}^{N_{t+1}}\mathsf{g}(\theta^{t+1},X^{t+1,i};\upsilon^t) + \frac{1}{N_{t+1}}\sum_{i=1}^{N_{t+1}}\mathsf{g}(\mathsf{T}(\upsilon^t),X^{t+1,i};\upsilon^t) - \mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t)
\end{aligned}$$

where we used that, by definition of θ^{t+1},

$$\frac{1}{N_{t+1}}\sum_{i=1}^{N_{t+1}}\mathsf{g}(\theta^{t+1},X^{t+1,i};\upsilon^t) - \frac{1}{N_{t+1}}\sum_{i=1}^{N_{t+1}}\mathsf{g}(\mathsf{T}(\upsilon^t),X^{t+1,i};\upsilon^t) \leq 0.$$

Since $\mathsf{G}(\theta^{t+1};\upsilon^t)-\mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t)\geq 0$ by definition of $\mathsf{T}(\upsilon^t)$, then

$$\left|\mathsf{G}(\theta^{t+1};\upsilon^t)-\mathsf{G}(\mathsf{T}(\upsilon^t);\upsilon^t)\right| \leq 2\sup_{(\theta,\upsilon)\in\mathbb{T}\times\mathbb{U}}\left|\frac{1}{N_{t+1}}\sum_{i=1}^{N_{t+1}}\mathsf{g}(\theta,X^{t+1,i};\upsilon)-\mathsf{G}(\theta;\upsilon)\right|.$$

The proof is concluded by application of Theorem 13 with $u\leftarrow(\theta,\upsilon)$, $\mathbb{V}\leftarrow\mathbb{T}\times\mathbb{U}$ and $\mathsf{h}(u,X)\leftarrow\mathsf{g}(\theta,X;\upsilon)$. Note indeed that A3-a implies B1-ii and B1-iii; A4 implies B1-iv; and A2-Item b implies B1-v.

6.6.2 Proof of Item ii and Item iii.

Lemma 16. *Let $\mathcal{K}$ be a compact subset of $\mathbb{T}$. The set $\mathbb{V}:=\{(\theta,\upsilon):\theta\in\mathcal{K},\upsilon\in\mathbb{U}[\theta]\}$ is a compact subset of $\mathbb{T}\times\mathbb{U}$.*

Proof. $\mathbb{V}$ is bounded as a subset of the compact set $\mathbb{T}\times\mathbb{U}$. Let us prove it is closed. Let $\{\theta^t,t\geq 0\}$ be a $\mathcal{K}$-valued sequence converging to $\theta^\star\in\mathcal{K}$, and let $\{\upsilon^t,t\geq 0\}$ be a sequence converging to $\upsilon^\star\in\mathbb{U}$ and such that $\upsilon^t\in\mathbb{U}[\theta^t]$; we prove that $\upsilon^\star\in\mathbb{U}[\theta^\star]$. Since F and G are continuous functions (see A3-c and Proposition 11), we have

$$\mathsf{F}(\theta^\star)=\lim_t\mathsf{F}(\theta^t)=\lim_t\mathsf{G}(\theta^t;\upsilon^t)=\mathsf{G}(\theta^\star;\upsilon^\star),$$

thus showing that $\upsilon^\star\in\mathbb{U}[\theta^\star]$. This concludes the proof. □

We apply Theorem 5 with $\mathcal{L}\leftarrow\mathcal{L}^+$.

Step 1. We first prove that $\mathcal{L}^+$ is closed. Let $\{\theta^t,t\geq 0\}$ be a $\mathcal{L}^+$-valued sequence, converging to $\theta^\star$; let us show that $\theta^\star\in\mathcal{L}^+$.
There exists a sequence $\{\upsilon^t,t\geq 0\}$ such that $\upsilon^t\in\mathbb{U}[\theta^t]$ and $\mathsf{T}(\upsilon^t)=\theta^t$ for all $t\geq 0$. Since $\{\upsilon^t,t\geq 0\}$ is a $\mathbb{U}$-valued sequence, there exists $\upsilon^\star\in\mathbb{U}$ and a subsequence $\{\upsilon^{\rho(t)},t\geq 0\}$ such that $\lim_t\upsilon^{\rho(t)}=\upsilon^\star$. It holds

$$\mathsf{F}(\theta^\star)=\lim_t\mathsf{F}(\theta^{\rho(t)})=\lim_t\mathsf{G}(\theta^{\rho(t)};\upsilon^{\rho(t)})=\mathsf{G}(\lim_t\theta^{\rho(t)};\lim_t\upsilon^{\rho(t)})=\mathsf{G}(\theta^\star;\upsilon^\star)$$

where we used that F and G are continuous on $\mathbb{T}$ and $\mathbb{T}\times\mathbb{U}$ respectively (see A3-c and Proposition 11). In addition,

$$\theta^\star=\lim_t\theta^{\rho(t)}=\lim_t\mathsf{T}(\upsilon^{\rho(t)})=\mathsf{T}(\lim_t\upsilon^{\rho(t)})=\mathsf{T}(\upsilon^\star),$$

where we used that T is continuous on $\mathbb{U}$ (see Proposition 12). Hence, there exists $\upsilon^\star\in\mathbb{U}[\theta^\star]$ such that $\mathsf{T}(\upsilon^\star)=\theta^\star$: $\theta^\star\in\mathcal{L}^+$.

Step 2. By A3-b and A3-c, T is a point-to-point map and F is continuous.

Step 3. The condition H-i follows from A1 (see (7)).

Step 4. Let $\mathcal{K}$ be a compact subset of $\mathbb{T} \setminus \mathcal{L}^+$ and set $\mathbb{V} := \{(\theta, \upsilon) : \theta \in \mathcal{K}, \upsilon \in \mathbb{U}[\theta]\}$. By Lemma 16, $\mathbb{V}$ is a compact subset of $\mathbb{T} \times \mathbb{U}$. For $(\theta, \upsilon) \in \mathbb{V}$, $\mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\upsilon)) > 0$ since $\theta \notin \mathcal{L}^+$ (see (7) and A3-b). In addition, the function $(\theta, \upsilon) \mapsto \mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\upsilon))$ is continuous on $\mathbb{V}$. Hence, $\inf_{(\theta,\upsilon)\in\mathbb{V}} \mathsf{F}(\theta) - \mathsf{F}(\mathsf{T}(\upsilon)) > 0$ and this concludes the proof of H-ii.

Step 5. By Proposition 7, H-iii holds as soon as $\lim_t \mathsf{G}(\theta^{t+1}; \upsilon^t) - \mathsf{G}(\mathsf{T}(\upsilon^t); \upsilon^t) = 0$ and the condition (9) holds. The first limit holds true by Item i. Let us prove (9). Applying Lemma 16 with $\mathcal{K} := \mathbb{T}$ shows that $\{(\theta, \upsilon) : \theta \in \mathbb{T}, \upsilon \in \mathbb{U}[\theta]\}$ is a compact subset of $\mathbb{T} \times \mathbb{U}$. Therefore, since T is continuous on $\mathbb{U}$, the set $\mathcal{K}_\delta := \{(\theta, \theta') \in \mathbb{T} \times \mathbb{T}, \upsilon \in \mathbb{U}[\theta] : \|\theta' - \mathsf{T}(\upsilon)\| \geq \delta\}$ is compact. On $\mathcal{K}_\delta$, A3-b implies that $\mathsf{G}(\theta'; \upsilon) - \mathsf{G}(\mathsf{T}(\upsilon); \upsilon) > 0$. Since G and T are continuous, then the condition (9) is verified.

6.7 Technical results of Section 5

Lemma 17. *Let $q \in (0, 1)$. For all $y, u \in \mathbb{R}$, $|\rho_q(y + u) - \rho_q(y)| \leq |u|$.*

Proof. Set $\Delta := \rho_q(y+u) - \rho_q(y)$. Assume that $y + u \geq 0$ and $y \geq 0$. Then $\Delta = q(y+u) - qy = qu$ and $|\Delta| \leq |u|$.
Assume that $y + u < 0$ and $y < 0$. Then $\Delta = (q - 1)(y + u - y) = (q - 1)u$ and $|\Delta| \leq |u|$.
Assume that $y + u \geq 0$ and $y < 0$. This implies that $u > 0$ and $y \geq -u$. Then $\Delta = q(y + u) - (q - 1)y = y + qu \in [(q - 1)u, qu]$, and $|\Delta| \leq |u|$.
Assume that $y + u < 0$ and $y \geq 0$. This implies that $u < 0$ and $y < -u$. Then $\Delta = (q - 1)(y + u) - qy = -y + (q - 1)u \in [qu, (q - 1)u]$, and $|\Delta| \leq |u|$. □

Proof of Proposition 10 We write for any $\tau \in \mathbb{R}^\ell$

$$\left\langle \theta, \overline{W} \right\rangle = \left\langle \tau, \overline{W} \right\rangle + \frac{1}{\ell} \sum_{j=1}^{\ell} \Big(\ell(\theta_j - \tau_j) \overline{W}_j \Big),$$

and since the function $u \mapsto \rho_q(u)$ is convex, this yields

$$\rho_q\Big(Y - \left\langle \theta, \overline{W} \right\rangle\Big) \leq \frac{1}{\ell} \sum_{j=1}^{\ell} \rho_q\Big(Y - \left\langle \tau, \overline{W} \right\rangle - (\ell(\theta_j - \tau_j)\overline{W}_j)\Big).$$

Set

$$\mathsf{f}_\eta(\theta, X) := \rho_q(Y - \left\langle \theta, \overline{W} \right\rangle) - \rho_q(Y) + \eta\|\theta\|_1,$$

We obtain $\mathsf{f}_\eta(\theta, \cdot) \leq \mathsf{g}_\eta(\theta, \cdot; \tau)$ for any $\theta, \tau \in \mathbb{R}^\ell$. When $\tau = \theta$, it is easily checked that $\mathsf{g}_\eta(\theta, x; \theta) = \mathsf{f}_\eta(\theta, x)$. Since ρ_q is strictly convex, then $\mathbb{U}[\theta] = \{\theta\}$. This concludes the proof of A1. It is easily seen that $|\rho_q(Y + u) - \rho_q(Y)| \leq |u|$ (see e.g. Lemma 17) so that

$$|\mathsf{g}_\eta(\theta, X; \tau)| \leq (2\|\tau\| + \|\theta\|) \, \|\overline{W}\| + \eta\|\theta\|_1.$$

Hence, since $\mathbb{T}$ is a compact set, A3-a is satisfied with $p = p_\star$; remember that $\overline{W} := (1, W) \in \mathbb{R}^\ell$. A 3-b holds since ρ_q is strictly convex. Finally, since $u \mapsto \rho_q(u)$ is continuous, the functions $\theta \mapsto \mathsf{f}_\eta(\theta, x)$ and $(\theta, \tau) \mapsto \mathsf{g}_\eta(\theta, x; \tau)$ are continuous on $\mathbb{T}$ and $\mathbb{T} \times \mathbb{T}$ respectively, for all $x \in \mathbb{R} \times \mathbb{R}^\ell$. The proof of A3 is concluded by the dominated convergence Theorem.

The optimization step Let $q \in (0, 1)$ and $\eta > 0$. The `SAM2` update (5) in Algorithm 2, boils down to solving ℓ independent optimizations in 1D (see Proposition 10)

$$\text{argmin}_{u \in \mathbb{R}} \frac{1}{N_{t+1}} \sum_{i=1}^{N_{t+1}} \Big\{ \rho_q\Big(Y^{t+1,i} - \Big\langle \tau, \overline{W}^{t+1,i} \Big\rangle + \ell\, \overline{W}_j^{t+1,i} \tau_j - \ell \overline{W}_j^{t+1,i} u\Big) - \rho_q(Y^{t+1,i}) + \ell\eta|u| \Big\}.$$

Using notation $a_{N+1} := 0$, $b_{N+1} := 2\ell\eta N$ and $q_{N+1} = \frac{1}{2}$ to account for the penalty term, the 1D optimizations are all of the form

$$\arg\min_{u \in \mathbb{R}} g_{N,\eta}(u) \quad \text{with} \quad g_{N,\eta}(u) := \frac{1}{N} \sum_{i=1}^{N+1} \rho_{q_i}(a_i - b_i u); \tag{25}$$

where for all $i \in [N]$, $q_i = q$ and $(a_i, b_i) \in \mathbb{R}^2$ with b_i assumed different from 0 without loss of generality. For all $i \in [N+1]$, set $\mu_i := a_i/b_i$ with $\mu_{N+1} = 0$ and denote by $\mu_{(i,N+1)}$ the order numbers $\mu_{(1,N+1)} \le \mu_{(2,N+1)} \le \ldots \le \mu_{(N+1,N+1)}$. By definition of ρ_{q_i}, it holds

$$\begin{aligned} u \le \mu_i &\Longrightarrow \rho_{q_i}(a_i - b_i u) = |b_i|(\mu_i - u)\left(q_i 1_{b_i > 0} + (1 - q_i) 1_{b_i < 0}\right) \\ u \ge \mu_i &\Longrightarrow \rho_{q_i}(a_i - b_i u) = |b_i|(u - \mu_i)\left((1 - q_i) 1_{b_i > 0} + q_i 1_{b_i < 0}\right). \end{aligned}$$

This implies that the function $u \mapsto g_{N,\eta}(u)$ tends to $+\infty$ when $|u| \to +\infty$, is continuous on $\mathbb{R}$, and is linear on the intervals $\big(\mu_{(i,N+1)}, \mu_{(i+1,N+1)}\big]$ for all $i \in [N]$: a minimizer of $u \mapsto g_{N,\eta}(u)$ is $\text{argmin}_{i \in [N+1]} g_{N,\eta}(\mu_i)$. If $\eta = 0$, the same conclusion holds straightforwardly.

7 Conclusion

We have proposed a new `SAM2` algorithm that extends the applicability of MM algorithms in a stochastic optimization context where the objective majorizers cannot be observed or computed exactly but can be estimated through stochastic simulations. When compared to gradient approaches, MM algorithms are interesting because their progress toward the target does not critically depend on the tuning of hyperparameters such as step-sizes. In a stochastic context, combining MM with sample averaging, `SAM2` uses Monte-Carlo approximations of the majorizers constructed from samples of data. The setting of step-sizes is then replaced by successive samples of increasing sizes, whose exact values are not critical as illustrated numerically in our experiments. Another advantage of `SAM2` is that no smoothness assumptions are made on the objective and majorizers,

which covers a much larger number of situations than gradient-based approaches. Nevertheless, when compared to gradient update, the sample averaging update links successive estimates only through a warm-start-like relationship that may result in a more variable sequence of estimations. One standard way to reduce this variability is to use Polyak averaging. With our framework, another possibility is to add a Bregman term between two successive estimates. Such addition leads to a minor change in the majorizers definition while penalizing departure from the previous parameter value and producing much smoother sequences. Similarly sparsity constraints though a L_1 norm could easily be added without changing the applicability of `SAM2`. In practice, this could be of great practical interest to handle datasets both large in size and dimension.

References

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.

Andrews, D. W. K. (1992). Generic uniform convergence. *Econometric Theory*, 8, 241–257.

Atchadé, Y. F., Fort, G., & Moulines, E. (2017). On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10), 1–33.

Audet, C. & Hare, W. (2017). Derivative-free and blackbox optimization. *Springer Series in Operations Research and Financial Engineering*.

Bauschke, H. & Combettes, P. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York.

Beck, A. & Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3), 167–175.

Beck, A. & Teboulle, M. (2009). *Gradient-based algorithms with applications to signal-recovery problems*, (pp. 42–88). Cambridge University Press.

Bickel, P. J. & Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics*, volume 1. CRC Press.

Bohning, D. & Lindsay, B. R. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Mathematical Statistics*, 40, 641–663.

Bonnans, J. F. (2019). *Convex and stochastic optimization*. Springer.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4), 231–357.

Byrd, R. H., Hansen, S. L., Nocedal, J., & Singer, Y. (2016). A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2), 1008–1031.

Byrne, C. L. (2014). *Iterative optimization in inverse problems*. CRC Press.

Cadoni, S., Chouzenoux, E., Pesquet, J.-C., & Chaux, C. (2016). A block parallel majorize-minimize memory gradient algorithm. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 3194–3198).: IEEE.

Cappé, O. & Moulines, E. (2009). On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society B*, 71, 593–613.

Celeux, G. & Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastic Reports*, 41, 119–134.

Chalvidal, M., Chouzenoux, E., Fest, J.-B., & Lefort, C. (2023). Block delayed majorize-minimize subspace algorithm for large scale image restoration. *Inverse Problems*, 39(4), 044002.

Chen, G. & Teboulle, M. (1993). Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3), 538–543.

Chen, X., Liu, W., Mao, X., & Yang, Z. (2020). Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21, 1–43.

Chen, X., Liu, W., & Zhang, Y. (2019). Quantile regression under memory constraint. *Annals of Statistics*, 47, 3244–3273.

Chouzenoux, E. & Fest, J.-B. (2022). Sabrina: A stochastic subspace majorization-minimization algorithm. *Journal of Optimization Theory and Applications*, 195(3), 919–952.

Chouzenoux, E. & Pesquet, J.-C. (2016). Convergence rate analysis of the majorize–minimize subspace algorithm. *IEEE Signal Processing Letters*, 23(9), 1284–1288.

Combettes, P. L. & Pesquet, J.-C. (2015). Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2), 1221–1248.

Combettes, P. L. & Wajs, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4), 1168–1200.

Cui, Y. & Pang, J. (2022). *Modern Nonconvex Nondifferentiable Optimization*. Philadelphia: SIAM.

Dacunha-Castelle, D. & Duflo, M. (1986). *Probability and Statistics: Volume II*. Springer.

Davidson, J. (2021). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27, 94–128.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Dieuleveut, A., Fort, G., Moulines, E., & Robin, G. (2021). Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34, 29553–29566.

D'Orazio, R., Loizou, N., Laradji, I. H., & Mitliagkas, I. (2023). Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *Transactions on Machine Learning Research*.

Dragomir, R. A., Even, M., & Hendrikx, H. (2021). Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (pp. 2815–2825).: PMLR.

Fort, G., Gach, P., & Moulines, E. (2021a). Fast incremental expectation maximization for finite-sum optimization: nonasymptotic convergence. *Statistics and Computing*, 31, 1–24.

Fort, G. & Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4), 1220–1259.

Fort, G. & Moulines, E. (2023). Stochastic variable metric proximal gradient with variance reduction for non-convex composite optimization. *Statistics and Computing*, 33.

Fort, G., Moulines, E., & Wai, H.-T. (2021b). Geom-spider-em: Faster variance reduced stochastic expectation maximization for nonconvex finite-sum optimization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3135–3139).: IEEE.

Gourieroux, C. & Monfort, A. (1995). *Statistics And Econometrics Volume 2: Testing, Confidence Regions, Model Selection, And Asymptotic Theory*. Cambridge: Cambridge University Press.

Hall, P. & Heyde, C. (1980). *Martingale Limit Theory and Its Application*. Elsevier Science.

Hong, M., Wang, X., Razaviyayn, M., & Luo, Z.-Q. (2017). Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163(1), 85–114.

Hunter, D. R. & Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 9, 60–77.

Hunter, D. R. & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58, 30–37.

Ichinose, T., Yukawa, M., & Cavalcante, R. L. G. (2023). Online kernel-based quantile regression using Huberized pinball loss. In *2023 31st European Signal Processing Conference (EUSIPCO)* (pp. 1803–1807).

Jiang, R. & Yu, K. (2022). Renewable quantile regression for streaming data sets. *Neurocomput.*, 508(C), 208–224.

Karimi, B. & Li, P. (2021). Two-timescale stochastic em algorithms. In *2021 IEEE International Symposium on Information Theory (ISIT)* (pp. 890–895).: IEEE.

Karimi, B., Miasojedow, B., Moulines, E., & Wai, H.-T. (2019a). Non-asymptotic analysis of biased stochastic approximation scheme. *Proceedings of Machine Learning Research*, 99, 1–31.

Karimi, B., Wai, H.-T., Moulines, E., & Li, P. (2022). Minimization by incremental stochastic surrogate optimization for large scale nonconvex problems. In *International Conference on Algorithmic Learning Theory* (pp. 606–637).: PMLR.

Karimi, B., Wai, H.-T., Moulines, R., & Lavielle, M. (2019b). On the global convergence of (fast) incremental expectation maximization methods. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*.

Khanh Hien, L. T., Phan, D. N., Gillis, N., Ahookhosh, M., & Patrinos, P. (2022). Block Bregman majorization minimization with extrapolation. *SIAM Journal on Mathematics of Data Science*, 4(1), 1–25.

Lai, T. (1989). Extended stochastic liapounov functions and recursive algorithms in linear stochastic systems. In K. H. N. Christopeit & N. Kohlmann (Eds.), *Stochastic Differential Systems: Proceedings of the 4th Bad Honnef Conference, June, 20–24, 1988* (pp. 206–220).: Springer New-York.

Lan, G. (2020). *First-Order and Stochastic Optimization Methods for Machine Learning*. Springer Nature Switzerland AG.

Lan, G., Nemirovski, A., & Shapiro, A. (2012). Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134, 425–458.

Lange, K. (2013). *Optimization*. New York: Springer.

Lange, K. (2016). *MM Optimization Algorithms*. Philadelphia: SIAM.

Lange, K., Won, J.-H., Landeros, A., & Zhou, H. (2021). *Nonconvex Optimization via MM Algorithms: Convergence Theory*, (pp. 1–22). John Wiley & Sons, Ltd.

Lei, Y. & Zhou, D.-X. (2020). Convergence of online mirror descent. *Applied and Computational Harmonic Analysis*, 48(1), 343–373.

Lin, Z., Li, H., & Fang, C. (2020). Accelerated optimization for machine learning. *Nature Singapore: Springer*.

Lions, P. L. & Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6), 964–979.

Liu, A., Lau, V. K., & Kananian, B. (2019). Stochastic successive convex approximation for non-convex constrained stochastic optimization. *IEEE Transactions on Signal Processing*, 67(16), 4189–4203.

Liu, A., Lau, V. K., & Zhao, M.-J. (2018). Online successive convex approximation for two-stage stochastic nonconvex optimization. *IEEE Transactions on Signal Processing*, 66(22), 5941–5955.

Liu, J., Cui, Y., & Pang, J.-S. (2022). Solving nonsmooth and nonconvex compound stochastic programs with applications to risk measure minimization. *Mathematics of Operations Research*, 47(4), 3051–3083.

Liu, J. & Pang, J.-S. (2023). Risk-based robust statistical learning by stochastic difference-of-convex value-function optimization. *Operations Research*, 71(2), 397–414.

Lupu, D. & Necoara, I. (2023). Convergence analysis of stochastic higher-order majorization–minimization algorithms. *Optimization Methods and Software*, (pp. 1–30).

Mairal, J. (2013). Stochastic majorization-minimization algorithm for large-scale optimization. In *Advances in Neural Information Processing Systems* (pp. 2283–2291).

Mairal, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal of Optimization*, 25, 829–855.

McLachlan, G. J. & Krishnan, T. (2008). *The EM Algorithm And Extensions*. New York: Wiley.

McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Meng, X.-L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80, 267–278.

Meyn, S. (2022). *Control systems and reinforcement learning*. Cambridge University Press.

Mokhtari, A. & Koppel, A. (2020). High-dimensional nonconvex stochastic optimization by doubly stochastic successive convex approximation. *IEEE Transactions on Signal Processing*, 68, 6287–6302.

Naderi, S., He, K., Aghajani, R., Sclaroff, S., & Felzenszwalb, P. (2019). Generalized majorization-minimization. In *International Conference on Machine Learning* (pp. 5022–5031).: PMLR.

Nedić, A. & Lee, S. (2014). On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization*, 24(1), 84–107.

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

Nemirovskii, A. & Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley.

Nguyen, H. D., Forbes, F., Fort, G., & Cappé, O. (2022). An online minorization-maximization algorithm. In *Conference of the International Federation of Classification Societies* (pp. 263–271).: Springer.

Nitanda, A. (2014). Stochastic proximal gradient descent with acceleration techniques. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, volume 27: Curran Associates, Inc.

Polyak, R. A. (2021). *Introduction to Continuous Optimization*. Springer Cham.

Razaviyayn, M., Hong, M., & Luo, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal of Optimization*, 23, 1126–1153.

Razaviyayn, M., Sanjabi, M., & Luo, Z. (2016). A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming Series B*, (pp. 515–545).

Robbins, H. & Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi (Ed.), *Optimizing Methods in Statistics* (pp. 233–257). Academic Press.

Rosasco, L., Villa, S., & Vü, B. C. (2020). Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, 82.

Rossignol, C., Sureau, F., Chouzenoux, É., Comtat, C., & Pesquet, J.-C. (2022). A bregman majorization-minimization framework for pet image reconstruction. In *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 1736–1740).: IEEE.

Schraudolph, N. N., Yu, J., & Günter, S. (2007). A stochastic quasi-newton method for online convex optimization. In *Artificial intelligence and statistics* (pp. 436–443).: PMLR.

Serfling, R. J. (1980). *Approximation Theorems Of Mathematical Statistics*. New York: Wiley.

Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2021). *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica*, 15, 831–840.

van der Vaart, A. & Wellner, J. (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Vidyasagar, M. (2003). *Learning and generalisation: with applications to neural networks*. London: Springer.

Yuille, A. L. & Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, 15, 915–936.

Zhang, H., Zhou, P., Yang, Y., & Feng, J. (2019). Generalized majorization-minimization for non-convex optimization. In *IJCAI* (pp. 4257–4263).

Zhang, S. & He, N. (2018). *On the Convergence Rate of Stochastic Mirror Descent for Nonsmooth Nonconvex Optimization*. Technical report, arXiv 1806.04781.

Zheng, S. (2011). Gradient descent algorithms for quantile regression with smooth approximation. *International Journal of Machine Learning and Cybernetics*, 2(3), 191–207.

Ziegler, K. (2001). Uniform laws of large numbers for triangular arrays of function-indexed processes under random entropy conditions. *Results in Mathematics*, 39, 374–389.