Central European University
Department of Mathematics and its Applications

Kinga Tikosi

Convergence results regarding stochastic gradient descent methods for dependent data streams

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics

Supervisor: Miklós Rásonyi

2021
Budapest, Hungary

Declaration

I, Kinga Tikosi, candidate for the degree of Doctor of Philosophy at the Central European University Department of Mathematics and its Applications, declare herewith that the present thesis is based on my research and only such external information as properly credited in notes and bibliography.

I declare that no unidentified and illegitimate use was made of the work of others, and no part of this thesis infringes on any person's or institution's copyright. I also declare that no part of this thesis has been submitted to any other institution of higher education for an academic degree.

Budapest, Hungary, September 2021 _____

Contents

List of Figures

List of Tables

Preface

The idea of gradient descent dates back to [Cauchy et al., 1847] and it is one of the most well-known optimization methods. It is also very intuitive: the negative of the gradient describing which way the function descends the fastest, one just starts at an arbitrary point and takes steps in the direction of the negative of the gradient to reach a point where there is no further descent: a local minimum.

The topic of this thesis is to study the convergence of stochastic gradient methods. In practice, we might not be in a situation where the gradient is easily accessible and this is the source of the stochasticity. This can be due to noisy measurements or insufficient information.

This thesis is organized as follows. Chapter 1 is an introduction to recursive approximation and stochastic gradient methods in general as well as the summary of important results about these algorithms.

Chapter 2 is devoted to applications, here we describe how the different versions of stochastic gradient descent are used in machine learning, and in various applications from mathematical finance.

Chapter 3 is based on the preprint [Rásonyi and Tikosi, 2020] and studies the convergence of the Kiefer-Wolfowitz algorithm. This algorithm was designed to maximize a real function $U(\theta)$, which is only available through noisy measurements. This means that we observe another function $J(\theta, X)$ that is an unbiased estimate of $U(\theta)$, i.e. $\mathbb{E}J(\theta, X) = U(\theta)$, where X represents the noise. Starting from some initial guess θ_0 the recursive algorithm reads

$$\theta_{n+1} = \theta_n + a_n H(\theta_n, X_{n+1}, c_n),$$

where $H(\theta_n, X_{n+1}, c_n)$ plays the role of the gradient estimate being a difference quotient of two noise-corrupted measurements defined as

$$H(\theta_n, X_{n+1}, c_n) = \frac{J(\theta + c_n, X^1_{n+1}) - J(\theta - c_n, X^2_{n+1})}{2c_n},$$

with appropriately chosen deterministic sequences (a_n) and (c_n). The variables X^1 and X^2 were assumed to be independent in each step, however in our analysis we do not

assume that. In Therorem 3.2.1 we prove that under suitable technical assumptions, the Kiefer-Wolfowitz algorithm has a convergence rate $O(n^{-1/5})$ in mean absolute error, while Theorem 3.2.2 shows that the fixed gain version of the algorithm will track the optimum. The novelty of these results is that we do not assume differentiability, not even continuity of $\theta \to J(\theta, \cdot)$ and the sequence X_n is not assumed to be i.i.d., it may well be dependent as long as it satisfies a mixing condition.

Chapter 4 is based on the preprint [Rásonyi and Tikosi, 2021] and studies the convergence of the stochastic gradient Langevin algorithm. Starting from some initial guess θ_0^λ the recursion defined as

$$\theta_{n+1}^\lambda = \theta_n^\lambda - \lambda H(\theta_n^\lambda, Y_n) + \sqrt{\lambda}\xi_{n+1}, n \in \mathbb{N},$$

where $H(\theta, Y_n)$ is an unbiased estimate of $\nabla U(\theta)$ and $\lambda \in (0, 1]$. This recursion was designed to sample from the distribution proportional to $e^{-U(\theta)}$ as it is the discretization of the Langevin SDE, with ∇U replaced by $H(\theta_t^\lambda, Y_t)$. In the main result Therorem 4.2.1 we prove that under suitable assumptions, the law of θ_n converges to a limiting law in total variation distance. The novelty of this result is once again that we do not assume that the noise sequence is i.i.d.

All the simulations and numeric experiments, as well as the illustrations for this work were made in Python. While the code is not included here, the author is happy to share it upon inquiry.

Chapter 1

Stochastic gradient methods

Stochastic approximation methods (often abbreviated SA) are iterative procedures for solving optimization problems. In this chapter we present the classical results in this area that the rest of the thesis builds on.

The earliest and most well-known stochastic iterative schemes are the Robbins-Monro (RM) algorithm [Robbins and Monro, 1951] and the Kiefer-Wolfowitz algorithm (KW) [Kiefer and Wolfowitz, 1952]. These algorithms can be regarded as stochastic versions of the classical gradient descent for optimization (therefore often referred to as stochastic gradient descent or SGD, an abbreviation we will use in the rest of this work as an umbrella term for the below presented schemes). In the context of these algorithms the stochasticity is due to noise-corrupted observations or insufficient information.

1.1 Deterministic recursive approximation

The zero-search of a continuous function as well as the classical gradient descent are well-known methods in numerics for recursive approximation, see e.g. [Duflo, 2013, Polyak, 1987]. We recall them to put the stochastic methods in context.

The following recursion will converge to a point where a continuous function crosses a given level.

Theorem 1.1.1 *(Zero search of a deterministic continuous function) Suppose that f is a continuous real function such that $f(\theta^*) = \alpha$ and such that, for all θ, $(f(\theta) - \alpha)(\theta - \theta^*) < 0$ and $|f(\theta)| \leq K(1 + |\theta|)$ for some constant K. Suppose that (a_n) is a positive sequence such that $\sum a_n = \infty$ and $a_n \to 0$. Then the sequence (θ_n) defined by*

$$\theta_{n+1} = \theta_n - a_n (\alpha - f(\theta_n))$$

converges to θ^*, for all initial values θ_0.

Proof: Follows from Proposition 1.2.3 of [Duflo, 2013]. □

Remark 1.1.1 The special case when $a_n = f'(\theta_n)$ and $\alpha = 0$ is known as the Newton method (or Newton-Raphson algorithm). Newton's method has quadratic convergence in general (see Theorem 1 of Section 1.5.2 of [Polyak, 1987]), but it can be poorer in some cases (when the starting point is not well-chosen or when the function is not continuously differentiable).

For the search of the optimum of a differentiable function one can use gradient descent.

Theorem 1.1.2 *(Gradient descent) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable function, and that its gradient ∇f is Lipschitz continuous with constant $L > 0$. Then the iterates (θ_n) of gradient descent started from some deterministic initial value θ_0, defined by*

$$\theta_{n+1} = \theta_n - a_n \nabla f(\theta_n),$$

with a constant step size of $a_n = a \leq 1/L$ satisfy

$$|f(\theta^n) - f(\theta^*)| \leq C \frac{\|\theta_0 - \theta^*\|_2^2}{n}$$

where $f(\theta^*)$ is the optimal value and C is a positive constant depending on L.

Proof: The proof is straightforward estimations and uses telescoping sums. □

Clearly, if f is globally convex, this method will find the global optimum, otherwise the recursion might be caught up in some local minima.

Remark 1.1.2 Theorem 1.1.2 shows that gradient descent has a convergence rate of $O(1/n)$ for convex functions. This rate can be improved by having stricter assumptions (like strong convexity) or using a momentum term.

1.2 Robbins-Monro algorithm

The Robbins-Monro algorithm [Robbins and Monro, 1951] was introduced as a root-finding recursion, where the function is given in the form of an expected value. As seen below, the recursive algorithm (1.4) is a generalization of the classical gradient descent method, but instead of the gradient we have a stochastic quantity: an unbiased estimate of the gradient.

Assume that the function $h(\theta)$ is given and there is a unique θ^* such that $h(\theta^*) = \alpha$. The goal is to successively approximate the root of this equation, however, the classical

methods like the Newton-method cannot be used as the function $h(\theta)$ is not available, only an unbiased estimate, i.e. a function $H(\theta, X)$ with $h(\theta) = \mathbb{E}H(\theta, X)$ where X is a random variable. The authors proposed the recursion

$$\theta_{n+1} = \theta_n - a_{n+1}(\alpha - H(\theta_n, X_{n+1})), \tag{1.1}$$

where θ_0 is an arbitrary constant starting point, $(a_k)_{k\geq 1}$ is a sequence of positive real numbers such that

$$\sum_{k=0}^{\infty} a_k = \infty \tag{1.2}$$

$$\sum_{k=0}^{\infty} a_k^2 < \infty. \tag{1.3}$$

Robbins and Monro proved convergence in mean-square error and in probability provided that the function is non-decreasing, $h'(\theta^*)$ exists and is positive and $H(\theta, X)$ is uniformly bounded. Almost sure convergence of the algorithm is well known for Lipschitz functions, whenever the driving noise is a square integrable martingale difference sequence, see [Bhatnagar et al., 2013]. The proof relies on the so called ODE-method [Ljung, 1977] and [Ljung, 1980] and assumes global asymptotic stability of the ordinary differential equation associated with the problem. The speed of convergence in expectation is $O(n^{-1})$ when h is strongly convex and twice continuously differentiable and θ^* belongs to the interior of a convex set, see [Chung, 1954]. However with general convexity and without the smoothness the convergence rate is $O(n^{-\frac{1}{2}})$ only and it was proven in [Nemirovskij and Yudin, 1983] that it cannot be further improved.

Clearly, the RM method can be transformed to a minimum (or maximum) search if we treat the function h as the gradient of an objective function. Then the method translates to finding the local extrema of a function, when we can observe noisy measurements of its gradient. In this case we are looking for a θ^* such that $h(\theta^*) = 0$, altering the recursion to

$$\theta_{n+1} = \theta_n - a_n H(\theta_n, X_{n+1}). \tag{1.4}$$

In the rest of this thesis we will always refer to this form.

Remark 1.2.1 1. The sequence $(a_k)_{k\geq 1}$ is often called the **step size**, as it describes the size of the step one makes towards the gradient-estimate. Assumptions 1.2 and 1.3 on the step size intuitively say that while it needs to diminish, divergence of the sum is needed, i.e. the recursion needs to be have large enough steps to reach the optimum no matter how far away we start from it.

2. While the two assumptions 1.2, 1.3 are necessary for the convergence results, in practice sometimes constant step size is used, often referred to as **fixed gain stochastic approximation** as opposed to **decreasing gain**. With constant step size convergence is not guaranteed, but the recursion is meant to **track** the optimum if it is time-varying. Fixed gain recursion has been studied in the literature starting with [Kushner and Huang, 1981, Benveniste and Ruget, 1982] or more recently results in [Chau et al., 2019a] (which are relevant to our results).

3. In case of deterministic gradient descent as stated in Theorem 1.1.2 constant step size yields convergence since the gradient itself will converge to 0 as the iterations approach the optimum.

Remark 1.2.2 SGD vs. Monte Carlo: When unbiased estimates of the gradient are available, it can be a natural idea to use Monte Carlo method for optimization i.e. take a large number of observations, use their average as an estimate that is closer to the true gradient and perform classical, non-stochastic gradient descent. While this is a sensible strategy, SGD has the advantage of being an **online** method, which means that it always considers the new measurements and therefore it can adapt to slow changes in the dynamics too. A common practice in applications is to settle for a middle step between the two options and use a **mini-batch method**, i.e. take a number of measurements and use their mean as the gradient estimate:

$$\overline{H}(\theta_n, X_{n+1}) = \frac{1}{k}\sum_{i=1}^{k} H(\theta_n, X_{n+1}^k).$$

We will not go into discussing mini-batch methods in this thesis, but note that $\overline{H}$ is still an unbiased estimate with reduced variance, therefore the same analysis would apply.

1.3 Kiefer-Wolfowitz algorithm

The Kiefer-Wolfowitz algorithm [Kiefer and Wolfowitz, 1952] is often also referred to as finite differences stochastic approximation (or FDSA), as the main idea is to use a difference quotient as the gradient estimate. Here, the goal is to maximize (or minimize) an objective function $U : \mathbb{R}^d \to \mathbb{R}$, which is unknown, but one can observe the function U at any level, however, the observations are noise corrupted. In the original paper by Kiefer and Wolfowitz only the one dimensional version was introduced, for the multi-dimensional case that we present below, one can consult [Bhatnagar et al., 2013]. Comparing it to the Robbins-Monro algorithm, where noisy measurements of

the gradient of the function were taken, now we can take measurements of the function itself and therefore we will substitute the gradient-estimate $H(\theta_n, X_{n+1})$ in (1.4) with a finite difference of the form

$$H(\theta_n, X_{n+1}, c_n) = \sum_{i=1}^{d} \frac{J(\theta + c_n \mathbf{e}_i, X_{n+1}^1) - J(\theta - c_n \mathbf{e}_i, X_{n+1}^2)}{2c_n} \mathbf{e}_i,$$

where $J : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is measurable, X_{n+1}^1 and X_{n+1}^2 are random variables and $\mathbb{E}[J(\theta, X_0^{1,2})] = U(\theta),\ \theta \in \mathbb{R}^d$. The recursion therefore is

$$\theta_{n+1} = \theta_n + a_n H(\theta_n, X_{n+1}, c_n), \tag{1.5}$$

where the usual assumptions about the deterministic step size sequences $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ are:

$$\begin{aligned} c_k &\to 0,\ k \to \infty, \\ \sum_{k=1}^{\infty} a_k &= \infty, \\ \sum_{k=1}^{\infty} a_k c_k &< \infty, \\ \sum_{k=1}^{\infty} a_k^2 c_k^{-2} &< \infty. \end{aligned}$$

From the assumptions it follows that also $a_k \to 0$ as $k \to \infty$. A standard choice for these sequences that fulfills these assumptions is $a_k = ak^{-1}$ and $c_k = ck^{-\gamma}$, where $a > 0, c > 0$ and $\gamma \in (0, 1/2)$.

Remark 1.3.1

1. Intuitively, the first and third assumption on the step sizes are needed so that the gradient estimates become more precise, referring back to Remark 1.2.1, the divergence of the sequence (a_k) is still needed, while the last assumption requires (a_k) to converge to 0 quicker than (c_k), which ensures that the steps $a_{n+1}H(\theta_n, X_{n+1}, c_n)$ do not blow up.

2. While in the case of the RM algorithm the gradient estimates were assumed to be unbiased, note that in the present case the difference quotients in general are not unbiased estimates of the gradient.

In their original paper Kiefer and Wolfowitz proved convergence in probability, later Blum [Blum et al., 1954] proved almost sure convergence.

Spall [Spall et al., 1992] introduced a version of KW method where random directions are used. This is the so called simultaneous perturbation stochastic approximation

(SPSA) method which then became very popular due to its computational simplicity. In one dimension it is the same iteration as the KW method, but in d dimensions the latter will use $2d$ measurements in each step, 2 for each coordinate, and SPSA will only use 2 measurements in one random direction independent of the dimension:

$$H(\theta_n, X_{n+1}, c_n) = \frac{J(\theta + c_n\Delta_n, X^1_{n+1}) - J(\theta - c_n\Delta_n, X^2_{n+1})}{2c_n}\mathbf{e}_i,$$

where Δ_n is a d-dimensional vector with i.i.d. coordinates with zero mean, a standard choice being symmetrical Bernoulli ± 1 random variables as they satisfy the moment conditions that were proposed. More about SPSA and the conditions for almost sure convergence and asymptotic normality are available in [Spall, 2005, Gerencsér et al., 2007].

1.4 Stochastic gradient Langevin dynamics

Stochastic gradient Langevin dynamics (often abbreviated SGLD) is an iterative scheme where additional noise is introduced to a standard stochastic gradient method with the purpose of allowing the recursion to escape local extrema and find the global one for non-convex objective functions. Proposed in [Welling and Teh, 2011], the SGLD algorithm was designed to perform Bayesian learning on large data sets, as a computationally less expensive alternative to Markov chain Monte Carlo (MCMC) methods. For more about SGLD see [Raginsky et al., 2017, Brosse et al., 2018].

Remark 1.4.1 While the SGD methods (RM and KW) presented in the previous sections are designed with the goal of finding the optimum θ^*, that is finding the mode of the posterior distribution, the objective of SGLD is to sample from the posterior distribution.

The origin of the idea from Bayesian learning is explained below, yet the main task for us boils down to sampling from a distribution

$$\pi(A) := \frac{\int_A \mathrm{e}^{-U(\theta)}\mathrm{d}\theta}{\int_{\mathbb{R}^d} \mathrm{e}^{-U(\theta)}\mathrm{d}\theta},$$

where A is a Borel-set of $\mathbb{R}^d$ and $U : \mathbb{R}^d \to \mathbb{R}_+$ is a continuously differentiable function.

Recall the Langevin stochastic differential equation:

$$\mathrm{d}\Theta_t = -\nabla U(\Theta_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \tag{1.6}$$

where W is a d-dimensional standard Brownian motion on a filtered probability space $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\right)$ with the usual conditions[1]. Under mild assumptions on the drift term $\nabla U(\Theta)$ the Langevin SDE has a unique invariant probability π (with respect to the d-dimensional Lebesgue measure) that is proportional to $\exp(-U(x))$, $x \in \mathbb{R}^d$.

Remark 1.4.2 Note that $\mathrm{d}\Theta_t = -\nabla U(\Theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$ and $\mathrm{d}\Theta_t = -\frac{1}{2}\nabla U(\Theta_t)\,\mathrm{d}t + \mathrm{d}W_t$ generate the same dynamics and both are referred to as the Langevin SDE.

The Euler-discretization[2] of (1.6) is a discrete-time Markov chain, named unadjusted[3] Langevin algorithm (or short ULA), defined by

$$\theta_{t+1} = \theta_t - \lambda_t \nabla U(\theta_t) + \sqrt{2\lambda_t}\xi_{t+1},$$

where ξ_{t+1} is Gaussian noise.

To use this in the context of Bayesian learning, following [Welling and Teh, 2011], let $\theta \in \mathbb{R}^d$ be a parameter and $X = \{x_i, i \in \{1, 2, \dots N\}\}$ a set of data points with a large N. Assume that θ has a prior density $\pi_0(\theta)$. Then SGLD is meant to sample from the posterior distribution with density $\pi(\theta) = \pi(\theta|x) \propto \pi_0(\theta)\prod_{i=1}^N p(x_i|\theta)$ using the recursion

$$\theta_{t+1} = \theta_t - \lambda_t \left(\nabla \log \pi_0(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_t)\right) + \sqrt{2\lambda_t}\xi_{t+1}, \tag{1.7}$$

where $p(x_{t_i}|\theta_t)$ is the likelihood of the data conditioned on θ_t and ξ_{t+1} is Gaussian noise. In [Teh et al., 2016] the authors prove weak convergence of the recursion to π under assumptions on the drift and the step size sequence, that converges to 0.

Algorithm (1.7) can be interpreted as a minibatch version of the Euler-discretization of (1.6), where the gradient $\nabla U(\Theta_t)$ is substituted by its unbiased estimate. Therefore it is a combination of a stochastic gradient method and the discretization of the Langevin equation, where the gradient-part forces the iteration to step towards high

[1] A filtered probability space $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P\right)$ is said to satisfy the usual conditions if $(\mathcal{F}_t)_{t\geq 0}$ is right-continuous and $\mathcal{N} \subset \mathcal{F}_0$ with $\mathcal{N} = \{X \subset \Omega \mid \exists Y \in \mathcal{F} : X \subset Y \wedge P(Y) = 0\}$.

[2] The Euler-discretization (or Euler–Maruyama method) of an SDE $dX_t = a(t, X_t)\,dt + b(t, X_t)\,dW_t$, $X_0 = c$, where W_t is the standard Brownian motion, on an interval $0 \leq t \leq b$ is a numerical method solving the initial value problem on a mesh $0 = t_0 < t_1 < \cdots < t_k = b$. The step sizes are $h_n = t_n - t_{n-1}$ and the corresponding values $x_0, \dots, x_k$ are defined as $x_0 = c$ and recursively $x_{n+1} = x_n + h_{n-1}a(t_n, x_n) + b(t_n, x_n)\left(W_{t_{n+1}} - W_{t_n}\right)$, where x_n is the intended approximation of X_{t_n}.

[3] The adjusted version is usually referred to as Metropolis adjusted Langevin algorithm (MALA) and it includes a Metropolis-Hastings-type accept-reject mechanism to correct for the discretization error.

probability areas while the noise term ensures that it will explore the whole parameter space.

Using the notation $H(\theta, X)$ for the unbiased estimate, such that $\mathbb{E}H(\theta, X) = \nabla U(\theta_t)$, the algorithm becomes

$$\theta_{t+1} = \theta_t - \lambda_t H(\theta_t, X_t) + \sqrt{2\lambda_t}\xi_{t+1}, \tag{1.8}$$

which in fact is a similar recursion to the RM algorithm (1.4) with the presence of the additional noise sequence (ξ_t).

For the fixed gain version of algorithm (1.8) (where $\lambda_t = \lambda$ for all t) the law of the iterates converge to a limiting distribution π_λ, which is in general different from π, but for small λ it is close to it. The sampling error of θ_t^λ has been thoroughly analysed in the literature: $d(\mathcal{L}(\theta_t^\lambda), \mu)$ has been estimated for various probability metrics d, see [Chau et al., 2021, Barkhagen et al., 2021, Raginsky et al., 2017, Brosse et al., 2018], it is of the order $\sqrt{\lambda}$.

Chapter 2

Applications of SGD

Stochastic approximation algorithms can be useful in practice when the data collected is subject to noise, or when a mathematical model for the problem is too complicated and therefore one tries to optimize the system by adjusting the parameters.

In this chapter we provide examples of application areas of recursive stochastic approximation schemes, with special focus on cases where assuming a dependent data stream is sensible, as well as examples where discontinuous stochastic versions of the objective functions naturally arise. First we briefly explain arguably the most important application of SGD: training models in machine learning. Then we present various applications from mathematical finance. Here we assume that the reader is familiar with the basic financial models, and we explain only the crucial concepts for the examples.

2.1 SGD in machine learning

Recursive methods based on stochastic gradient descent are widely used in machine learning and statistical learning including support vector machines, generative adversarial networks etc. Combining it with *backpropagation*, SGD is the standard way for training an artificial neural network, which is what we present below. As we will see, in this setting the optimization problem can be high dimensional (with a complicated network the number of parameters to train can easily go up to one million) and SGD will reduce the computational burden by making the iterations fast.

In the case of feed-forward neural networks for supervised learning, as discussed in [Bottou et al., 2018], one starts out with a $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ collection of examples, where $x_i \in \mathcal{X}$ are the features of the data and $y_i \in \mathcal{Y}$ are the labels. The goal is to design a prediction function $h(\cdot, w) : \mathcal{X} \to \mathcal{Y}$ using the collection of learning examples, that maps the feature set into the labels, where w is the collection of all the parameters

that are to be optimised. In deep neural networks the value of the prediction function h will be computed by applying successive transformation to the input vector $x_i \in \mathbb{R}^{d_0}$. A fully connected layer performs

$$x_i^{(j)} = s(W_j x_i^{(j-1)} + b_j) \in \mathbb{R}^{d_j},$$

where $x_i(0) = x_i$, the matrix $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ and vector $b_j \in \mathbb{R}^{d_j}$ contain the jth layer parameters, and s is a component-wise nonlinear activation function. Then the output vector $x_i^{(J)}$ is the prediction function value $h(x_i; w)$, where the parameter vector w collects all the parameters $\{(W_1, b_1), \ldots, (W_J, b_J)\}$ of the successive layers.

The training of the network is done by minimising the empirical risk of misclassification:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(h(x_i; w), y_i),$$

where l is a loss function quantifying how much the prediction matches the correct label. Often another quadratic regularization term is added to the minimization problem to avoid the parameters taking extreme values:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(h(x_i; w), y_i) + \kappa |w|^2$$

with some $\kappa > 0$. With supervised learning, often the set of labels is a discrete set (e.g. in the case of identifying a written character, classifying texts, etc. the extent of misclassification can not always be quantified), but the loss function l is usually chosen in a way that it is continuous (the loss function is typically chosen by experimentation).

The minimization problem is then solved multiple times for a given training set with various candidates for h and the ultimate solution w^* is the one that yields the best performance on the validation set[1]. For this optimisation SGD is the standard method. The candidates for h might include various architectures of neural networks with different hyperparameters, e.g. different numbers and types of layers, different numbers of nodes in the layers and different parameters for the regularization.

The algorithm works as follows:

[1]The data set is typically divided into three subsets: the training set, the validation set and the testing set. One uses the training set to train all the models, then chooses the one which has the best performance on the validation set. Then the testing set is where the performance is checked. The testing set is different from the validation set, because, intuitively, we do not want to test our prediction function on the same exact dataset that made us choose that specific function.

Initialize w (randomly or e.g. to be the zero-vector).
Repeat for $t \in (1, \ldots, T)$:

1. Draw a random sample (x^t, y^t) from the training set with replacement.
2. Derive the prediction function $h(x^t, w)$.
3. Compute $\frac{\partial l(h(x^t;w),y^t)}{\partial w}$ by the chain rule layer by layer, this algorithm is called backpropagation.
4. Update $w \to w - \lambda_t \frac{\partial l(h(x^t;w),y^t)}{\partial w}$.

Therefore here the stochasticity comes from the random selection of the pair (x_i, y_i) from the learning set. In the machine learning setting, the step size is often called the **learning rate**. Backpropagation (short for backward propagation of errors) means that we calculate the derivative backwards: first the gradient of the final layer is computed, then we proceed layer by layer using the rules of differentiation. For this step it is convenient if the activation functions and the loss function are differentiable.

Referring back to 1.2.2, this method can be interpreted as an algorithm that uses only one measurement at a time, in contrast to the mini-batch methods, where the updating function is the average a few measurements, or the Monte-Carlo method, where the goal would be to use a large number of measurements (or all of the data points from the given data set, also known as batch gradient descent). Using mini-batch methods or SGD have the advantage of being online methods: further data points can be added anytime to improve the result without the need of recalculating anything.

2.2 Trading with a mean reverting asset

An application of stochastic approximation was presented in [Zhuang, 2008]. A mean-reverting process is considered as a model for the asset prices and the goal is to numerically develop a *buy-low-and-sell-high* strategy, i.e. to approximate two optimal thresholds for trading: the investor will buy the stock whenever the low price limit is hit, and sell when the price exceeds the high limit. In [Zhang and Zhang, 2008] a closed form is presented to solve such a problem, stochastic approximation therefore aims to reduce computational effort.

Let $X(t)$ be a a mean reverting process described by

$$dX(t) = a(b - X(t))dt + \sigma dW(t),\ X(0) = x,$$

where $a > 0$, $\sigma > 0$, $b \in \mathbb{R}$ and $W(t)$ is a standard Brownian motion. Let the stock price be given by the equation

$$S(t) = \exp(X(t)).$$

The buying and selling times are defined as hitting times

$$0 \leq \tau^{b_1} \leq \tau^{s_1} \leq \tau^{b_2} \leq \tau^{s_2} \leq \dots,$$

where τ^{b_1} is the first buying time, τ^{s_1} is the first selling time etc. A proportional transaction fee $0 < K < 1$ and a discount factor $\rho > 0$ are taken into consideration. Thus the goal is to find

$$\arg\max U(\theta), \text{ for } \theta = (\theta^1, \theta^2)^T \in (0,\infty) \times (0,\infty), \text{ where}$$
$$U(\theta) = \mathbb{E}\left[\sum_{i=1}^{\infty} \exp(-\rho\tau^{s_i}) S(\tau_i^s)(1-K) - \exp(-\rho\tau^{b_i}) S(\tau^{b_i})(1+K)\right].$$

Assume that the process $S(t)$ can be observed and then the stopping times can be computed as

$$\begin{aligned}
\tau_n^{b_1} &= \inf\{t > 0, S(t) \leq \exp(\theta_n^1)\},\\
\tau_n^{s_i} &= \inf\{t > \tau_n^{b_i}, S(t) \geq \exp(\theta_n^2)\}, \text{ for } i \geq 1,\\
\tau_n^{b_i} &= \inf\{t > \tau_n^{s_i - 1}, S(t) \leq \exp(\theta_n^1)\}, \text{ for } i \geq 2.
\end{aligned}$$

Combine the random effects in the vector

$$\xi_n = (S(\tau_n^{b_1}), S(\tau_n^{s_1}), \tau_n^{b_2}), S(\tau_n^{s_2}), \dots, \tau_n^{b_1}, \tau_n^{s_1}, \dots).$$

To design the recursive algorithm, gradient estimates with averaged samples are used, however similar results can be achieved without averaging, so we will stick to presenting only the latter version as it fits better in the present context.

Assume that a function $J(\theta, \xi)$ can be observed and it fulfills $\mathbb{E}J(\theta, \xi) = U(\theta)$ for each θ. Let $J_n^{\pm} = (J_n^{\pm,1}, J_n^{\pm,2})^T$ be two measurements from the simulations, defined by

$$J_n^{\pm,i}(\theta, \xi_n^{\pm}) = J(\theta \pm \delta_n e_i, \xi_n^{\pm}), \text{ for } i = 1, 2,$$

where e_i are the standard unit vectors, $\delta_n \to 0$ is the difference sequence, and ξ^+ and ξ^- are two different aggregate noise process taken respectively at the threshold values $\theta \pm \delta_n e_i$. Then the gradient estimate is defined as $H(\theta_n, \xi_n) = \frac{J_n^+ - J_n^-}{2\delta_n}$ and the stochastic scheme has the form

$$\theta_{n+1} = \theta_n + \varepsilon_n H(\theta_n, \xi_n),$$

where $(\varepsilon_n)_{n \in \mathbb{N}}$ is a sequence of positive step sizes.

Remark 2.2.1 In practice, this would mean that having the data sequence of historical prices, the algorithm uses the utility of the income generated by trading at the stopping times given by the actual threshold values over a rolling window of training data, using the newest data points to make a number of updates. The algorithm would work as follows:

Initialize θ ideally within a couple of percents of the current price as we want the stock prices to hit these thresholds eventually.
Repeat:

1. Collect historical stock price data for the last M time points (this can be daily or even minute price data)
2. Repeat for N learning steps on this data: Compute $\sum_{i=1}^{\infty} \exp(-\rho\tau^{s_i})S(\tau_i^s)(1-K) - \exp(-\rho\tau^{b_i})S(\tau^{b_i})(1+K)$ for $\theta \pm \delta_n e_i$ and update θ based on these measurements (in this case there will of course be a finite number of transactions).
3. Wait until m new data points are available, where $m < M$.

Note that although concrete dynamics of the system are assumed, when using the algorithm on real data, one does not need to estimate the parameters describing the dynamics. On the other hand if the parameters describing the dynamics are known, they can be used to generate (independent) simulations to play the role of $J(\theta, \xi_n^{\pm})$.

In [Zhuang, 2008] weak convergence of the scheme is proven under assumptions on the sequences (δ_n) and (ε_n), continuity of the second derivative $U_{\theta\theta}(.)$, and appropriate integrability and mixing conditions.

2.3 Stock liquidation

In [Yin et al., 2002] the authors proposed a finite differences recursive algorithm for stock liquidation, which was one of the first applications of stochastic approximation in financial mathematics. When trading with a stock, a crucial question is when to close one's position and that is to be determined by a selling rule. Such a selling rule is a pair of threshold levels: trading happens once the *target price* or the *stop-loss limit* (i.e. a threshold where the trader closes the position in order to prevent further loss) was hit. They consider a so called *switching Black-Scholes model* defined as follows.

Let $\alpha(t)$ be a finite-state continuous time Markov chain, where each member of the states-space $\mathcal{M} = \{1, \ldots, m\}$ describes a specific economic environment. The easiest

such example would be $m = 2$ with $\alpha(t) = 1$ meaning economic growth with an upward trend and $\alpha(t) = 2$ meaning a downward trend. The stock price $S(t)$ satisfies the SDE

$$\frac{dS(t)}{S(t)} = \mu(\alpha(t))dt + \sigma(\alpha(t))dW(t), \tag{2.1}$$

where $W(.)$ with a standard Brownian motion independent of $\alpha(.)$ and the initial price $S(0) = S_0$. When the underlying Markov-chain has more than two states, the optimal solution can be difficult to obtain in a closed form as a solution of a two-boundary problem, although it is known that such a solution exists. A stochastic approximation algorithm therefore aims to reduce computational effort in such a case. Defining the process

$$X(t) = \int_0^t r(\alpha(s))ds + \sigma(\alpha(s))dW(s), \tag{2.2}$$

with $r(i) = \mu(i) - \frac{\sigma^2(i)}{2}$ for all $i \in \mathcal{M}$ one can rewrite the solution of (2.1) as $S(t) = S_0 \exp(X(t))$. The pair of thresholds are defined as $\theta = (\theta^1, \theta^2) \in \mathbb{R} \times \mathbb{R}$, the stock is liquidated as soon as the price process reaches $S_0 \exp(-\theta^1)$ or $S_0 \exp(\theta^2)$.

The optimisation task is to find the optimal threshold values so that the expected return is maximal. Define the stopping time

$$\tau = \inf\{t > 0 : S(t) \notin (S_0 \exp(-\theta^1), S_0 \exp(\theta^2))\}.$$

Then the goal is to find

$$\arg\max U(\theta), \text{ for } \theta \in \mathbb{R} \times \mathbb{R},$$
$$U(\theta) = \mathbb{E}[J(X(\tau)) \exp(-\rho\tau)],$$

where $\Phi : \mathbb{R} \to \mathbb{R}$ is a suitable utility function and ρ is the discount rate.

Similarly to the previous section, we will only discuss the algorithm without averaging.

To get a gradient estimate they propose the following method: use (2.2) to generate sample paths of $X(t)$. At the nth iteration the threshold values are $\theta_n = (\theta_n^1, \theta_n^2)^T$ and the corresponding stopping time τ_n can be computed. Use the notation $\xi_n = (X(\tau_n), \tau_n)^T$ to combine the random effects from of the process $X(t)$ and the stopping time.

Let $J(\theta, \xi)$ be a real-valued function and assume that $\mathbb{E}J(\theta, \xi_n^{\pm}) = U(\theta)$ for each θ. Let $J_n^{\pm} = (J_n^{\pm,1}, J_n^{\pm,2})^T$ be two (not necessarily independent) measurements, defined by

$$J_n^{\pm,i}(\theta, \xi_n^{\pm}) = J(\theta \pm \delta_n e_i, \xi_n^{\pm}), \text{ for } i = 1, 2,$$

where e_i are the standard unit vectors and ξ^+ and ξ^- are two different aggregate noise processes taken respectively at the threshold values $\theta \pm \delta_n e_i$. Then the gradient estimate is defined as $H = \frac{J_n^+ - J_n^-}{2\delta_n}$ and the stochastic scheme has the form

$$\theta_{n+1} = \theta_n + \varepsilon_n H(\theta_n, \xi_n),$$

where $(\varepsilon_n)_{n\in\mathbb{N}}$ is a sequence of positive step sizes.

In practice we can use the same noisy measurements as in Remark 2.2.1.

In [Yin et al., 2002] weak convergence of the scheme is proven under assumptions on the sequences (δ_n) and (ε_n), continuity of the second derivative $U_{\theta\theta}(.)$, and appropriate integrability and mixing conditions. They also demonstrate the use of the algorithm on real market data.

2.4 Trailing stop

A similar recursive algorithm was presented in [Zhuang, 2008] for trailing stop orders. A trailing stop order, just like a stop-loss order is a tool to limit an investor's loss in the case of the stock price dropping, however this time the aim is to adjust to the market price and initiate the selling at a given percentage below the market price. Thus if the market price increases, the stop price also increases, whereas if the price drops, the stop price does not drop and the stock is liquidated as soon as it is hit. Let $S(t)$ be an observable stochastic process modelling the stock price and $h \in (0, 1)$ be the trailing stop percentage. Then at time t the stop price is defined as

$$T_h(t) = (1 - h)S_{max}(t),$$

where S_{max} is the maximum price observed, i.e. $S_{max}(t) = \max\{S(u) : 0 \leq u \leq t\}$. Define the stopping time $\tau = \inf\{t > 0 : S(t) \leq T_h(t)\}$. The goal is to find

$$\arg\max U(h), \text{ for } h \in [a, 1], \text{ where}$$
$$U(h) = \mathbb{E}\left[\Phi(S(\tau)) \exp(-\rho\tau)\right],$$

where $a > 0$ is a lower bound for the trailing stop percentage, ρ is the discount rate and $\Phi(S) = \frac{S - S_0}{S_0}$ is the reward function. A stochastic approximation algorithm is proposed in [Zhuang, 2008] for this problem, with the same design as the previous examples.

2.5 Recursive computation of V@R and CV@R

Value at risk (V@R) and conditional value at risk (CV@R, also known as expected shortfall) are two widely used risk measures in finance, which aim to quantify the tail

risk of a portfolio or a financial instrument. In what follows we present a recursive scheme to compute V@R and CV@R that can be found in [Laruelle and Pagès, 2012] as an application to the convergence theorem established by the authors. The recursive method to compute V@R and CV@R using stochastic approximation was introduced in [Bardou et al., 2009].

Let X be a random variable representing the loss of a financial instrument.

Definition 2.5.1 The V@R at level $\alpha \in (0,1)$ of a portfolio is the α-quantile of the distribution X, i.e.

$$V@R_\alpha(X) = \inf\{\theta : \mathbb{P}(X \leq \theta) \geq \alpha\}.$$

Assume that X has a positive continuous density function f_X on $\mathbb{R}$. Then $\theta_\alpha = V@R_\alpha(X)$ is the unique solution of

$$\mathbb{P}(X > \theta_\alpha) = 1 - \alpha.$$

In application sometimes CV@R is preferred, since as opposed to V@R, it is a coherent risk measure, meaning it satisfies the subadditivity property.

Definition 2.5.2 Let $X \in L^1(\mathbb{P})$ have an atomless distribution. The CV@R at level $\alpha \in (0,1)$ is the conditional expectation of loss assuming that the loss exceeds the V@R at level α:

$$CV@R_\alpha(X) = \mathbb{E}[X | X \geq V@R_\alpha(X)].$$

The SA scheme for V@R and CV@R relies on a formulation of these risk measures as the solutions to an optimization problem introduced in [Uryasev and Rockafellar, 2001].

Proposition 2.5.1 Let $X \in L^1(\mathbb{P})$ have an atomless distribution. Then the function $V(\theta) = \theta + \frac{1}{1-\alpha}E(X-\theta)^+, \theta \in \mathbb{R}$ is convex, and

$$CV@R_\alpha(X) = \min_\theta \left(\theta + \frac{1}{1-\alpha}E(X-\theta)^+\right),$$
$$V@R_\alpha(X) = \inf \arg\min_\theta \left(\theta + \frac{1}{1-\alpha}E(X-\theta)^+\right).$$

Proof: See [Uryasev and Rockafellar, 2001]. □

Then defining $H(\theta, y) = 1 - \frac{1}{1-\alpha}\mathbb{1}_{\{y \geq \theta\}}$ yields $V'(\theta) = \mathbb{E}H(\theta, X)$ and one can solve the optimization problem and thus find the V@R with the stochastic gradient scheme

$$\theta_{k+1} = \theta_k - \lambda_k H(\theta_n, X_n), n \geq 0.$$

Note that one could solve the problem with classical gradient descent $\theta_{k+1} = \theta_n - \frac{\lambda_k}{1-\alpha}\left(F_X(\theta_k) - \alpha\right)$, but in general the cdf might not have a closed form or it might be computationally demanding in higher dimensions. For the computation of CV@R the following algorithm can be used with $\zeta_0 = 0$ and $v(\theta, x) = \theta + \frac{(x-\theta)^+}{1-\alpha}$:

$$\zeta_{n+1} = \zeta_n - \frac{1}{n+1}(\zeta_n - v(\theta_n, X_n)), n \geq 0.$$

In [Laruelle and Pagès, 2012] the authors prove almost sure convergence of the scheme under appropriate assumption about the noise, smoothness and Lyapunov assumptions and suitable (λ_k) stepsizes.

2.6 Implicit correlation search

The following application is presented in [Laruelle and Pagès, 2012] as well.

As the model for the financial market a two-dimensional Black-Scholes model is considered with one riskless asset $X_t^0 = e^{rt}$ and two risky assets

$$X_t^i = x_0^i \exp(r - \frac{\sigma_i^2}{2})t + \sigma_i W_t^i, i = 1, 2,$$

for $t \geq 0$, with initial price $x_0^i > 0$, $i = 1, 2$, where $\langle W^1, W^2 \rangle_t = \rho t, \rho \in [-1, 1]$. A so called *best-of call* option with strike price K on these assets is a basket option with the payoff $(\max(X_T^1, X_T^2) - K)_+$.

Thus the aim of the stochastic approximation procedure is to solve the equation

$$P_{BoC}(x_0^1, x_0^2, K, \sigma_1, \sigma_2, r, \rho, T) = P_0^{market}$$

for ρ, where P_0^{market} is the price observed on the market and

$$\begin{aligned} &P_{BoC}(x_0^1, x_0^2, K, \sigma_1, \sigma_2, r, \rho, T) \\ &\quad = e^{-rT}\mathbb{E}\left[(\max(X_T^1, X_T^2) - K)_+\right] \\ &\quad = e^{-rT}\mathbb{E}\left[\left(\max\left(x_0^1 e^{\mu_1 T + \sigma_1\sqrt{T}Z^1}, x_0^2 e^{\mu_2 T + \sigma_2\sqrt{T}(\rho Z^1 + \sqrt{1-\rho^2})Z^2}\right) - K\right)_+\right], \end{aligned}$$

where $\mu_i = r - \frac{\sigma_i^2}{2}, i = 1, 2, Z = (Z^1, Z^2) \sim N(0, I_2)$. It is assumed that the equation has at least one solution (say ρ^*) and a trigonometric parametrization $\rho = \cos\theta, \theta \in \mathbb{R}$ is used as $\rho \in [-1, 1]$. For simplicty denote $P(\theta) = P_{BoC}(x_0^1, x_0^2, K, \sigma_1, \sigma_2, r, \cos(\theta), T)$. Then finding the implicit correlation translates to solving $P(\theta) = P_0^{market}$. Set

$$H(\theta, z) = e^{-rT}\left(\max\left(x_0^1 e^{\mu_1 T + \sigma_1\sqrt{T}z^1}, x_0^2 e^{\mu_2 T + \sigma_2\sqrt{T}(z^1\cos\theta + z^2\sin\theta)}\right) - K\right)_+ - P_0^{market},$$

for $\theta \in \mathbb{R}$ and $z \in \mathbb{R}^2$. Then $h(\theta) = \mathbb{E}H(\theta, Z_1) = P(\theta) - P_0^{market}$ and the recursive algorithm is defined for $n \geq 0$ as

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, Z_{n+1}), \text{ with } Z_{n+1} \sim N(0, I_2).$$

The authors once again prove almost sure convergence under appropriate assumptions.

Chapter 3

Convergence of the Kiefer-Wolfowitz algorithm in the presence of discontinuities

This chapter is based on the preprint [Rásonyi and Tikosi, 2020]. In what follows we estimate the expected error of the Kiefer-Wolfowitz stochastic approximation algorithm where the maximum of a function is found using finite differences of a stochastic representation of that function. An error estimate of the order $n^{-1/5}$ for the nth iteration is achieved using suitable parameters. The novelty with respect to previous studies is that we allow the stochastic representation to be discontinuous and to consist of possibly dependent random variables (satisfying a mixing condition).

3.1 Introduction

We are interested in maximizing a function $U : \mathbb{R}^d \to \mathbb{R}$ which is unknown. However, we can observe a sequence $J(\theta, X_n)$, $n \geq 1$ where $J : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is measurable,

$$\mathbb{E}[J(\theta, X_1)] = U(\theta), \ \theta \in \mathbb{R}^d, \tag{3.1}$$

and X_n, $n \geq 1$ is an $\mathbb{R}^m$-valued stationary process in the strong sense [1]. The stochastic representations $J(\theta, X_n)$ are often interpreted as noisy measurements of $U(\theta)$. In this paper we focus on applications to mathematical finance, described in Section 3.6 below, where $J(\theta, X_t)$ are functionals of observed economic variables X_t and θ determines an investor's portfolio strategy. In that context, stochasticity does not come from

[1] The process $(x_k)_{k\in\mathbb{Z}}$ is called strongly stationary (or strictly stationary) if the distribution is time invariant, i.e. the joint distribution of $(x_{t_1}, \dots x_{t_k})$ is the same as of $(x_{t_1+j}, \dots x_{t_k} + j)$ for every $t_1, \dots, t_k$ indeces and for all k and j.

measurement errors but it is an intrinsic property of the system. Maximizing U serves to find the best investment policy in an online, adaptive manner.

We study a recursive algorithm employing finite differences, as proposed in [Kiefer and Wolfowitz, 1952]. This is a variant of the Robbins-Monro stochastic gradient method [Robbins and Monro, 1951] where, instead of the objective function itself, its gradient is assumed to admit a stochastic representation.

The novelty in our work is that we do not assume differentiability, not even continuity of $\theta \to J(\theta, \cdot)$ and the sequence X_n may well be dependent as long as it satisfies a mixing condition. The only result in such a setting that we are aware of is in [Laruelle and Pagès, 2012], however, they only study almost sure convergence, without a convergence rate. Our purpose is not to find the weakest possible hypotheses but to arouse keen interest in the given problem that can lead to further, more general results. Our work is also a continuation of [Fort et al., 2016, Chau et al., 2019a], where discontinuous stochastic gradient procedures were treated.

The main theorems are stated in Section 3.2 and proved in Section 3.3. Section 3.4 recalls earlier results that we are relying on. A numerical example is provided in Section 3.5. We explain the significance of our results for algorithmic trading in Section 3.6.

3.2 Setup and results

For real-valued quantities X, Y the notation $O(X) = Y$ means that there is a constant $C > 0$ such that $|X| \leq CY$. We will always work on a fixed probability space $(\Omega, \mathcal{F}, P)$ equipped with a filtration $\mathcal{F}_n$, $n \in \mathbb{N}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. A decreasing sequence of sigma-algebras $\mathcal{F}_n^+$, $n \in \mathbb{N}$ is also given, such that, for each $n \in \mathbb{N}$, $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent and X_n is adapted to $\mathcal{F}_n$. The notation $\mathbb{E}[X]$ refers to the expectation of a real-valued random variable X, while $\mathbb{E}_k[X]$ is a shorthand notation for $\mathbb{E}[X|\mathcal{F}_k]$, $k \in \mathbb{N}$. $P_k(A)$ refers to the conditional probability $P(A|\mathcal{F}_k)$. We denote by $\mathbb{1}_A$ the indicator of a set A. The notation ω refers to a generic element of Ω. For $r \geq 1$, we refer to the set of random variables with finite rth moments as L^r. $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^k$ where k may vary according to the context.

For $i = 1, \ldots, d$, let $\mathbf{e}_i \in \mathbb{R}^d$ denote the vector whose ith coordinate is 1 and the other coordinates are 0. For two vectors $v, w \in \mathbb{R}^m$ the relation $v \leq w$ expresses that $v^i \leq w^i$ for all the components $i = 1, \ldots, m$. Let $B_r := \{\theta \in \mathbb{R}^d : |\theta| \leq r\}$ denote the ball of radius r, for $r \geq 0$.

Let the function $U : \mathbb{R}^d \to \mathbb{R}$ have a unique maximum at the point $\theta^* \in \mathbb{R}^d$. Consider

the following recursive stochastic approximation scheme for finding θ^*:

$$\theta_{k+1} = \theta_k + \lambda_k H(\theta_k, X_{k+1}, c_k), \text{ for } k \in \mathbb{N}, \tag{3.2}$$

starting from some initial (deterministic) guess $\theta_0 \in \mathbb{R}^d$, where H is an estimator of the gradient of J, defined as

$$H(\theta, x, c) = \sum_{i=1}^{d} \frac{J(\theta + c\mathbf{e}_i, x) - J(\theta - c\mathbf{e}_i, x)}{2c} \mathbf{e}_i,$$

for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$ and $c > 0$.

The sequences $(\lambda_k)_{k\in\mathbb{N}}$ and $(c_k)_{k\in\mathbb{N}}$ appearing in (3.2) will consist of positive real numbers, which are to be specified later. We will distinguish the cases where λ_k, c_k tend to zero and where they are kept constant, the former being called *decreasing gain* approximation and the latter *fixed gain* approximation.

Remark 3.2.1 Our results below could easily be formulated in a more general setting where $J(\theta_k + c_k\mathbf{e}_i, X_{k+1}(i))$ and $J(\theta_k - c_k\mathbf{e}_i, X'_{k+1}(i))$, $i = 1, \ldots, d$ are considered with distinct $X_{k+1}(i)$ and $X'_{k+1}(i)$. In the applications that motivate us this is not the case hence, for reasons of simplicity, we stay in the present setting.

Assumption 3.2.1 *U is continuously differentiable with unique maximum $\theta^* \in \mathbb{R}^d$. Denote $G(\theta) = \nabla U(\theta)$. The function G is assumed Lipschitz-continuous with Lipschitz-constant L_G.*

We assume in the sequel that the function J in (3.1) has a specific form. Note that though J is not continuous, U can nonetheless be continuously differentiable, by the smoothing effect of randomness.

Assumption 3.2.2 *Let the function J be of the following specific form:*

$$J(\theta, x) = l_0(\theta)\mathbb{1}_{A_0(x)} + \sum_{i=1}^{m_s} \mathbb{1}_{A_i(x)} l_i(\theta, x),$$

where $l_i : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ are Lipschitz-continuous (in both variables) for $i = 1, \ldots, m_s$ and, for some $m_p, m'_p \in \mathbb{N}$,

$$A_i(x) := \left(\cap_{j=1}^{m_p}\{\theta : x \leq g_i^j(\theta)\}\right) \bigcap \left(\cap_{j=1}^{m'_p}\{\theta : x > h_i^j(\theta)\}\right), \ i = 1, \ldots, m_s$$

with Lipschitz-continuous functions $g_i^j, h_i^j : \mathbb{R}^d \to \mathbb{R}^m$. Furthermore, $A_0(x) := \mathbb{R}^d \setminus \cup_{i=1}^{m_s} A_i(x)$ and

$$\cup_{x\in\mathbb{R}^m} \cup_{i=1}^{m_s} A_i(x) \subset B_D$$

for some $D > 0$. The function l_0 is twice continuously differentiable and there are constants L_1'', L_2'' such that

$$L_1'' I \leq \nabla\nabla l_0 \leq L_2'' I$$

where I is the $d \times d$ identity matrix.

Remark 3.2.2 Assumption 3.2.2 implies that ∇l_0 grows linearly, hence l_0 itself is locally Lipschitz with linearly growing Lipschitz-coefficient, that is,

$$|l_0(\theta_1) - l_0(\theta_2)| \leq L_0(1 + |\theta_1| + |\theta_2|)|\theta_1 - \theta_2|,$$

with some $L_0 > 0$, for all $\theta_1, \theta_2 \in \mathbb{R}^d$.

In plain English, we consider J which is smooth on a finite number of bounded domains (the interior of the constraint sets $A_i(x)$, $i = 1, \ldots, m_s$) but may have discontinuities at the boundaries. Furthermore, J (and hence also U) is required to be quadratic "near infinity" (on $A_0(x)$).

We briefly explain why such a hypothesis is not restrictive for real-life applications. Normally, there is a compact set Q (e.g. a cube or a ball) such that only parameters from Q are relevant, i.e. U is defined only on Q. Assume it has some stochastic representation

$$U(\theta) = \mathbb{E}[J(\theta, X_0)], \ \theta \in Q \tag{3.3}$$

and a unique maximum $\theta^* \in Q$. Assume that $Q \subset B_D$ for some D. Extend U outside B_D as $U(\theta) = -A|\theta|^2 + B$ for suitable A, B. Extend U and J to $B_D \setminus Q$ as well in such a way that U is continuously differentiable, $U(\theta) < U(\theta^*)$ for all $\theta \neq \theta^*$ (see Section 4 of [Chau et al., 2019b] for a rigorous construction of this kind). Set $J := U$ outside Q. Then our maximization procedure can be applied to this setting for finding θ^*.

Defining $U = l_0$ (essentially) quadratic outside a compact set is one way of solving the problem that such procedures often leave their effective domain Q. Other solutions are resetting, see e.g. [Gerencsér, 1992]; or an analysis of the probability of divergence, see e.g. [Benveniste et al., 1990].

The next assumption postulates that the process X should be bounded and the conditional laws of X_{k+1} should be absolutely continuous with a bounded density.

Assumption 3.2.3 *For each $k \in \mathbb{N}$,*

$$P_k(X_{k+1} \in A) = \int_A p_k(u, \omega)\, du, \ \ a.s., A \in \mathcal{B}(\mathbb{R}^d),$$

for some measurable $p_k : \mathbb{R}^d \times \Omega \to \mathbb{R}_+$ and there is a fixed constant F such that $p_k(u, \omega) \leq F$ holds for all k, ω, u. The random variable X_0 satisfies $|X_0| \leq K_0$ for some constant K_0.

Note that, by strong stationarity, the process X_k is uniformly bounded under Assumption 3.2.3.

We will assume a certain mixing property about the process X_n which we recall now. A family of $\mathbb{R}^d$-valued random variables Z_i, $i \in \mathcal{I}$ is called *L^r-bounded* for some $r \geq 1$ if $\sup_{i\in\mathcal{I}} \mathbb{E}|Z_i|^r < \infty$, here $\mathcal{I}$ may be an arbitrary index set.

For a random field $Y_n(\theta)$, $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ bounded in L^r for some $r \geq 1$, we define, for all $n \in \mathbb{N}$,

$$M_r^n(Y) = \operatorname*{ess\,sup}_\theta \sup_{k\in\mathbb{N}} \mathbb{E}^{1/r}[|Y_{n+k}(\theta)|^r \,|\mathcal{F}_n],$$
$$\gamma_r^n(\tau, Y) = \operatorname*{ess\,sup}_\theta \sup_{k\geq\tau} \mathbb{E}^{1/r}\left[\left|Y_{n+k}(\theta) - \mathbb{E}[Y_{n+k}(\theta)|\mathcal{F}^+_{n+k-\tau} \vee \mathcal{F}_n]\right|^r |\mathcal{F}_n\right], \tau \geq 0,$$
$$\Gamma_r^n(Y) = \sum_{\tau=0}^{\infty} \gamma_r^n(\tau, Y).$$

These quantities clearly make sense also for any L^r-bounded stochastic process Y_n, $n \in \mathbb{N}$ (the essential suprema disappear in this case). $M_r^n(Y)$ measures the (conditional) moments of Y while $\Gamma_r^n(Y)$ describes its dependence structure (like covariance decay). In particular, one can define $M_r^n(X)$, $\Gamma_r^n(X)$. We clearly have $M_r^n(X) \leq K_0$ under Assumption 3.2.3. The quantities $\Gamma_r^n(X)$ will figure in certain estimates later.

Assumption 3.2.4 *For some $\epsilon > 0$, $\gamma_3^n(\tau, X) = O((1+\tau)^{-4-\epsilon})$, where the constant of $O(\cdot)$ is independent of ω, τ and n. Furthermore,*

$$\mathbb{E}\left[\left|X_{n+k} - \mathbb{E}[X_{n+k}|\mathcal{F}_n^+]\right|\right] = O(k^{-2-\epsilon}),\ k \geq 1,$$

where the constant of $O(\cdot)$ is independent of n, k.

Both requirements in Assumption 3.2.4 are about how the effect of the past on the present decreases as we go back farther in time.

Example 3.2.1 Let ε_n, $n \in \mathbb{N}$ be a bounded i.i.d. sequence in $\mathbb{R}^m$ with bounded density w.r.t. the Lebesgue measure and choose $\mathcal{F}_k := \sigma(\varepsilon_j,\ j \leq k)$ and $\mathcal{F}_k^+ := \sigma(\varepsilon_j,\ j \geq k+1)$. Then $X_n := \varepsilon_n$, $n \in \mathbb{N}$ satisfies Assumptions 3.2.3 and 3.2.4.

Example 3.2.2 A causal infinite moving average process whose coefficients decay sufficiently fast is another pertinent example. Indeed, using the argument of Lemma 4.2 of [Chau et al., 2019a] one can show that $X_n := \sum_{j=0}^{\infty} s_j \varepsilon_{n-j}$, $n \in \mathbb{N}$ satisfies Assumption 3.2.4 where the ε_i are as above, $s_0 \neq 0$ and $|s_j| \leq (1+j)^{-\beta}$ holds for some $\beta > 9/2$. Assumption 3.2.3 is also clearly satisfied in that model.

Remark 3.2.3 A random field $Y_n(\theta), n \in \mathbb{N}$ is called uniformly conditionally L-mixing if $Y_n(\theta)$ is adapted to the filtration $\mathcal{F}_n$, $n \in \mathbb{N}$ for all θ, and the sequences $M_r^n(Y)$, $\Gamma_r^n(Y)$, $n \in \mathbb{N}$ are bounded in L^r for each $r \geq 1$. Our Assumption 3.2.4 thus requires a sort of related mixing property.

The concept of conditional L-mixing was introduced in [Chau et al., 2019a], inspired by [Gerencsér, 1989].

3.2.1 Decreasing gain stochastic approximation

The usual assumption on the sequences $(\lambda_k)_{k=1,2,\dots}$ and $(c_k)_{k=1,2,\dots}$ in the definition of the recursive scheme (3.2) are the following, see [Kiefer and Wolfowitz, 1952]:

$$\begin{aligned} c_k &\to 0, \ k \to \infty, \\ \sum_{k=1}^{\infty} \lambda_k &= \infty, \\ \sum_{k=1}^{\infty} \lambda_k c_k &< \infty, \\ \sum_{k=1}^{\infty} \lambda_k^2 c_k^{-2} &< \infty. \end{aligned} \tag{3.4}$$

In the sequel we stick to a more concrete choice which clearly fulfills the conditions in (3.4) above.

Assumption 3.2.5 *We fix $\lambda_0, c_0 > 0$, $\gamma \in (0, 1/3)$ and set*

$$\lambda_k = \lambda_0 \int_k^{k+1} \frac{1}{u} \, du,$$

and $c_k = c_0 k^{-\gamma}$, $k \geq 1$. We also assume $c_0 \leq 1$.

Asymptotically λ_k behaves like λ_0/k. However, our choice somewhat simplifies the otherwise already involved theoretical analysis.

The ordinary differential equation associated with the problem is

$$\dot{y}_t = \frac{\lambda_0}{t} G(y_t). \tag{3.5}$$

The idea to use an associated deterministic ODE to study the asymptotic properties of recursive schemes was introduced in [Ljung, 1977]. The intuition behind this association is that on the long run the noise effects average out and the asymptotic behavior is determined by this 'mean' differential equation.

Remark 3.2.4 Heuristic connection between the dynamics of the recursive scheme and the ODE can be seen if one looks at the Euler-discretization[2] of the latter. Recall that the Euler discretization of ODE (3.5) with the initial value $y_0 = \theta_0$ and step size 1 is

$$y_{k+1} = y_k + \frac{\lambda_0}{k} G(y_k).$$

This gives the step sizes λ_k a different interpretation: the time-step in the ODE.

The solution of (3.5) with initial condition $y_s = \xi$ will be denoted by $y(t, s, \xi)$ for $0 < s \leq t$.

Assumption 3.2.6 *The ODE (3.5) fulfills the stability assumption formulated below: there exist $C^* > 0$ and $\alpha > 0$ such that*

$$\left| \frac{\partial y(t, s, \xi)}{\partial \xi} \right| \leq C^* \left(\frac{s}{t} \right)^{\alpha \lambda_0}$$

for all $0 < s < t$.

Our main result comes next.

Theorem 3.2.1 *Let Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5 and 3.2.6 hold. Then*

$$\mathbb{E}|\theta_n - \theta^*| = O(n^{-\chi} + n^{-\alpha}),\ n \geq 1,$$

where $\chi = \min\{\frac{1}{2} - \frac{3}{2}\gamma, \gamma\}$ and the constant in $O(\cdot)$ depends only on θ_0.

To get the best result set $\gamma = \frac{1}{5}$. In this case the convergence rate is $\chi = \frac{1}{5}$ (provided that $\alpha \geq 1/5$). For Kiefer-Wolfowitz procedures [Sacks, 1958] establishes a convergence rate $n^{-1/3}$ under fairly restrictive conditions (e.g. J is assumed smooth and X is i.i.d.). Our approach is entirely different from that of [Sacks, 1958] and relies on the ODE method (see e.g. [Kushner and Clark, 2012]) in the spirit of [Gerencsér, 1992,Gerencsér, 1999,Gerencsér, 1998] where so-called SPSA procedures were analysed.

Theoretical analysis in the present case is much more involved for two reasons: the discontinuities of J and the state-dependent setting (hardly analysed in the literature at all). Our results are closest to [Gerencsér, 1998] where a rate of $n^{-2/7}$ is obtained for the SPSA algorithm (a close relative of Kiefer-Wolfowitz) imposing strong smoothness assumptions on J. As already remarked, in the absence of smoothness ours is the first study providing a convergence rate. Eventual strengthening of our result seems to be difficult and will be object of further investigations.

[2]The Euler discretization (as described in [Ascher and Petzold, 1998]) of an ODE $y' = f(t, y)$, $y(0) = c$ on an interval for $0 \leq t \leq b$ is a first order numerical method solving the initial value problem on a mesh $0 = t_0 < t_1 < \cdots < t_k = b$. The step sizes are $h_n = t_n - t_{n-1}$ and the corresponding values $y_0, \ldots, y_k$ are defined as $y_0 = c$ and recursively $y_n = y_{n-1} + h_n f\left(t_{n-1}, y_{n-1}\right)$, where y_n is the intended approximation of $y(t_n)$.

3.2.2 Fixed gain stochastic approximation

Let us also consider a modified recursive scheme

$$\theta_{k+1} = \theta_k + aH(\theta_k, X_{k+1}, c), \ k \in \mathbb{N}, \tag{3.6}$$

where a and c are fixed (small) positive reals, independent of k. In contrast with the previous scheme (3.2), which is meant to converge to the maximum of the function, this method is expected to *track* the maximum.

The ordinary differential equations associated with the problem are

$$\dot{y}_t = aG(y_t), \tag{3.7}$$

for each $a > 0$.

Remark 3.2.5 Here we refer back to Remark 3.2.4, noting that the Euler dicretization of ODE (3.7) with the initial value $y_0 = \theta_0$ and step size 1 is

$$y_{k+1} = y_k + aG(y_k).$$

Lemma 3.2.1 *Assumption 3.2.6 on the ODE (3.5) implies (3.7) being exponentially stable, i.e. satisfying*

$$\left|\frac{\partial y(t,s,\xi)}{\partial \xi}\right| \leq C^* e^{-\alpha a(t-s)}, \ 0 < s \leq t$$

for some $\alpha > 0$ (possibly different from the one in (3.5)).

Proof: Let Assumption 3.2.6 hold for ODE (3.5). Use exponential time-change with $s = e^u$ and $t = e^v$, and note that in the fixed gain case a will take the role of λ_0. Then we have the reparametrized ODE

$$\dot{y}_t = \dot{y}_{e^v} = \frac{\lambda_0}{e^v} G(y_t) e^v = \lambda_0 G(y_t),$$

as well as the stability

$$\left|\frac{\partial y(t,s,\xi)}{\partial \xi}\right| = \left|\frac{\partial y(e^v,e^u,\xi)}{\partial \xi}\right| \leq C^* \left(\frac{e^u}{e^v}\right)^{\alpha\lambda_0} = C^* e^{-\alpha\lambda_0(t-s)}, \ 0 < s \leq t.$$

□

Theorem 3.2.2 *Let Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4 and 3.2.6 hold. Then*

$$\mathbb{E}|\theta_n - \theta^*| = O\left(\max\left(c^2, \sqrt{\frac{a}{c}}\right) + e^{-a\alpha n}\right)$$

holds for all $n \geq 1$ where the constant in $O(\cdot)$ depends only on θ_0.

Note that, similarly to the decreasing gain setting, this leads to the best choice being $c = a^{\frac{1}{5}}$. We know of no other papers where the fixed gain case has been treated. In the case of stochastic gradients there are many such studies obtaining a rate of $\sqrt{a}$ for step size a, see e.g. [Chau et al., 2019a] and the references therein.

3.3 Proofs

The following lemma will play a pivotal role in our estimates: it establishes the *conditional* Lipschitz-continuity of the difference function obtained from J.

Lemma 3.3.1 *Under Assumptions 3.2.2 and 3.2.3, there is $C_\flat > 0$ such that, for each $i = 1, \dots, d$ and $c \leq 1$,*

$$\begin{aligned}&\mathbb{E}_k|J(\bar\theta_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar\theta_1 - c\mathbf{e}_i, X_{k+1}) - J(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}) + J(\bar\theta_2 - c\mathbf{e}_i, X_{k+1})|\\ &\leq C_\flat[|\bar\theta_1 - \bar\theta_2| + c^2]\end{aligned}$$

holds for all $k \in \mathbb{N}$ and for all pairs of $\mathcal{F}_k$-measurable $\mathbb{R}^d$-valued random variables $\bar\theta_1$, $\bar\theta_2$.

Proof: We assume that $m_s = 1$, $m_p = 1$, $m_p' = 0$. We will shortly refer to the general case later. We thus assume that $J(\theta, x) = l_1(\theta, x)1_{\{x \leq g(\theta)\}} + l_0(\theta)1_{A_0(x)}$ with some Lipschitz-continuous g, l_1 with Lipschitz-constant L_1 (for both). Let K_1 be an upper bound for l_1 in B_{D+2}.

Consider first the event $A_1 := \{\bar\theta_1, \bar\theta_2 \in B_{D+1}\}$ and the corresponding indicator $I_1 := 1_{A_1}$. Note that on I_1 we have $\bar\theta_j \pm c\mathbf{e}_i \in B_{D+2}$, $j = 1, 2$. Now estimate

$$\begin{aligned}
&\mathbb{E}_k|I_1 l_1(\bar\theta_1 + c\mathbf{e}_i, X_{k+1})1_{\{X_{k+1} \leq g(\bar\theta_1 + c\mathbf{e}_i)\}} - I_1 l_1(\bar\theta_2 + c\mathbf{e}_i, X_{k+1})1_{\{X_{k+1} \leq g(\bar\theta_2 + c\mathbf{e}_i)\}}|\\
\leq\ &\mathbb{E}_k|I_1 l_1(\bar\theta_1 + c\mathbf{e}_i, X_{k+1})1_{\{X_{k+1} \leq g(\bar\theta_1 + c\mathbf{e}_i)\}} - I_1 l_1(\bar\theta_2 + c\mathbf{e}_i, X_{k+1})1_{\{X_{k+1} \leq g(\bar\theta_1 + c\mathbf{e}_i)\}}|\\
+\ &\mathbb{E}_k|I_1 l_1(\bar\theta_2 + c\mathbf{e}_i, X_{k+1})1_{\{X_{k+1} \leq g(\bar\theta_1 + c\mathbf{e}_i)\}} - I_1 l_1(\bar\theta_2 + c\mathbf{e}_i, X_{k+1})1_{\{X_{k+1} \leq g(\bar\theta_2 + c\mathbf{e}_i)\}}|\\
\leq\ &L_1\mathbb{E}_k|\bar\theta_1 - \bar\theta_2|\\
+\ &K_1\sum_{j=1}^m \left[P_k(g^j(\bar\theta_2 + c\mathbf{e}_i) < X_{k+1}^j \leq g^j(\bar\theta_1 + c\mathbf{e}_i)) + P_k(g^j(\bar\theta_1 + c\mathbf{e}_i) < X_{k+1}^j \leq g^j(\bar\theta_2 + c\mathbf{e}_i))\right]\\
\leq\ &L_1|\bar\theta_1 - \bar\theta_2| + 2mK_1L_1F|\bar\theta_1 - \bar\theta_2|.
\end{aligned}$$

In the same way, we also get

$$\mathbb{E}_k|I_1 l_1(\bar\theta_1 - c\mathbf{e}_i, X_{k+1}) - I_1 l_1(\bar\theta_2 - c\mathbf{e}_i, X_{k+1})| \ \leq\ L_1|\bar\theta_1 - \bar\theta_2| + 2mK_1L_1F|\bar\theta_1 - \bar\theta_2|.$$

As l_0 is clearly Lipschitz on B_{D+2}, we also have

$$|I_1 l_0(\bar\theta_1 \pm c\mathbf{e}_i, X_{k+1}) - I_1 l_0(\bar\theta_2 \pm c\mathbf{e}_i, X_{k+1})| = O(|\bar\theta_1 - \bar\theta_2|).$$

Let L_2'' be an upper bound for the second derivative $\nabla\nabla l_0$, recall Assumption 3.2.2. Now let A_2 be the event that the line from $\bar\theta_1$ to $\bar\theta_2$ does not intersect B_{D+1} at all, let

$I_2 := 1_{A_2}$. It follows in particular that neither $\bar{\theta}_1 \pm c\mathbf{e}_i$ nor $\bar{\theta}_2 \pm c\mathbf{e}_i$ fall into B_D. Since $J = l_0$ outside B_D we can write, by the Lagrange mean value theorem,

$$\begin{aligned}
& \mathbb{E}_k I_2 |J(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}) + J(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1})| \\
= \;& 2c\mathbb{E}_k I_2 |\partial_{\theta_i} l_0(\xi_1) - \partial_{\theta_i} l_0(\xi_2)| \\
\leq \;& 2c \sup_{u \in \mathbb{R}^d} |\nabla(\partial_{\theta_i} l_0(u))| \, \mathbb{E}_k |\xi_1 - \xi_2| \\
\leq \;& 2cL_2'' \mathbb{E}_k |\xi_1 - \xi_2| \\
\leq \;& 2cL_2''[|\bar{\theta}_1 - \bar{\theta}_2| + 2c] \leq 2L_2'' |\bar{\theta}_1 - \bar{\theta}_2| + 4c^2 L_2''
\end{aligned}$$

holds with some random variables $\xi_j \in [\bar{\theta}_j - c\mathbf{e}_i, \bar{\theta}_j + c\mathbf{e}_i]$, $j = 1, 2$, remembering our assumptions on l_0 and $c \leq 1$.

Turning to the event $\Omega \setminus (A_1 \cup A_2)$, let us consider the directed straight line from $\bar{\theta}_1(\omega)$ to $\bar{\theta}_2(\omega)$ and let its first intersection point with the boundary of B_{D+1} be denoted by $\kappa_1(\omega)$ and its second intersection point by $\kappa_2(\omega)$. In the case where there is only one intersection point it is denoted by $\kappa_1(\omega)$. Let I_3 be the indicator of the event that there is only one intersection point (κ_1) with B_{D+1} and that $\bar{\theta}_1$ is inside B_{D+1}. The arguments of the previous two cases guarantee that

$$\begin{aligned}
& \mathbb{E}_k I_3 |J(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}) + J(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1})| \\
\leq \;& \mathbb{E}_k I_3 |J(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}) - J(\kappa_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}) + J(\kappa_1 - c\mathbf{e}_i, X_{k+1})| \\
+ \;& \mathbb{E}_k I_3 |J(\kappa_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}) - J(\kappa_1 - c\mathbf{e}_i, X_{k+1}) + J(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1})| \\
= \;& O(|\bar{\theta}_1 - \kappa_1|) + O(|\kappa_1 - \bar{\theta}_2|) + O(c^2) \\
= \;& O(|\bar{\theta}_1 - \bar{\theta}_2|) + O(c^2).
\end{aligned}$$

Similarly, if I_4 is the indicator of the event where there is one intersection point and $\bar{\theta}_2$ is inside B_{D+1} then we also get

$$\begin{aligned}
& \mathbb{E}_k I_4 |J(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}) + J(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1})| \\
& \qquad = O(|\bar{\theta}_1 - \bar{\theta}_2| + c^2).
\end{aligned}$$

Let I_5 denote the indicator of the case where both $\bar{\theta}_1$, $\bar{\theta}_2$ are outside B_{D+1} and there are two intersection points κ_1, κ_2. We get, as above,

$$\begin{aligned}
& \mathbb{E}_k I_5 |J(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}) + J(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1})| \\
= \;& O(|\bar{\theta}_1 - \kappa_1|) + O(|\kappa_1 - \kappa_2|) + O(|\kappa_2 - \bar{\theta}_2|) + O(c^2) \\
= \;& O(|\bar{\theta}_1 - \bar{\theta}_2| + c^2).
\end{aligned}$$

Finally, in the remaining case (where there is only one intersection point with B_{D+1} though both $\bar{\theta}_1$, $\bar{\theta}_2$ are outside B_{D+1}) we similarly get an estimate of the order $O(|\bar{\theta}_1 - \bar{\theta}_2| + c^2)$ and hence we eventually obtain the statement of the lemma.

When $m_p = 0$ and $m'_p = 1$, the same ideas work. When $m_p + m'_p > 1$ we can rely on the following elementary observation:

$$\left| \prod_{j=1}^{m_p} \mathbb{1}_{\{X_{k+1} \leq g^j(\bar{\theta}_1 + c\mathbf{e}_i)\}} - \prod_{j=1}^{m_p} \mathbb{1}_{\{X_{k+1} \leq g^j(\bar{\theta}_2 + c\mathbf{e}_i)\}} \right| \leq \sum_{j=1}^{m_p} \left| \mathbb{1}_{\{X_{k+1} \leq g^j(\bar{\theta}_1 + c\mathbf{e}_i)\}} - \mathbb{1}_{\{X_{k+1} \leq g^j(\bar{\theta}_2 + c\mathbf{e}_i)\}} \right|,$$

and its counterpart for the h^j. Estimates can be repeated for each summand in the definition of J so the case $m_s > 1$ follows, too. □

The arguments of the previous lemma, (3.8) in particular, also give us the following:

Lemma 3.3.2 *Under Assumptions 3.2.2 and 3.2.3, there is $C_\natural > 0$ such that, for each $i = 1, \ldots, d$,*

$$\mathbb{E}_k |J(\bar{\theta} + c\mathbf{e}_i, X_{k+1}) - J(\bar{\theta} - c\mathbf{e}_i, X_{k+1})| \leq C_\natural c, \quad 0 < c \leq 1,$$

holds for all $k \in \mathbb{N}$ and for all $\mathcal{F}_k$-measurable B_D-valued random variables $\bar{\theta}$. □

3.3.1 Moment estimates

In this subsection, we will prove that the first moments of our iteration scheme remain bounded. We start with a preliminary lemma on deterministic sequences.

Lemma 3.3.3 *Let $x_k \geq 0$, $k \in \mathbb{N}$ be a sequence, let $\zeta_k > 0$, $k \geq 1$ be another sequence. If they satisfy $\nu\zeta_k < 1$, $k \geq 1$ and*

$$x_k \leq (1 - \nu\zeta_k)x_{k-1} + \underline{c}\zeta_k, \ \ k \geq 1,$$

with some $\underline{c}, \nu > 0$ then

$$\sup_{k \in \mathbb{N}} x_k \leq x_0 + \frac{\underline{c}}{\nu}.$$

Proof: Following the argument of Lemma 1 in [Durmus and Moulines, 2017], we notice that

$$x_k \leq \prod_{i=1}^{k}(1 - \nu\zeta_i)x_0 + \underline{c}\sum_{i=1}^{k}\zeta_i \prod_{j=i+1}^{k}(1 - \nu\zeta_j),$$

where an empty product is meant to be 1. We can write

$$\begin{aligned} & \sum_{i=1}^{k}\zeta_i \prod_{j=i+1}^{k}(1 - \nu\zeta_j) \\ = \ & \frac{1}{\nu}\sum_{i=1}^{k}\left(\prod_{j=i+1}^{k}(1 - \nu\zeta_j) - \prod_{j=i}^{k}(1 - \nu\zeta_j) \right) \\ \leq \ & \frac{1}{\nu}. \end{aligned}$$

This shows the claim. □

Certain calculations are easier to carry out if we consider the continuous time embedding of the discrete time processes. Consider the following extension θ_t, $t \in \mathbb{R}_+$ of θ_k, $k \in \mathbb{N}$: let

$$\theta_t := \theta_k + \int_k^t a_u H(u, \theta_k)\, du$$

for all $k \in \mathbb{N}$ and for all $k \leq t < k+1$, where $H(t, \theta) = H(\theta, X_{k+1}, c_k)$ for all $k \in \mathbb{N}$ and for all $k \leq t < k+1$, $c_t = c_k$ and $a_u = \lambda_0 / \max\{u, 1\}$, $u \geq 0$. Extend the filtration to continuous time by $\mathcal{F}_t := \mathcal{F}_{\lceil t \rceil}$, $t \in \mathbb{R}_+$. Now fix $\mu > 1$. We introduce an auxiliary process that will play a crucial role in later estimates. For each $n \geq 1$ and for $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$ define $\overline{y}_t := y(t, \lceil n^\mu \rceil, \theta_{\lceil n^\mu \rceil})$, i.e. the solution of (3.5) starting at $\lceil n^\mu \rceil$ with initial condition $\overline{y}_{\lceil n^\mu \rceil} = \theta_{\lceil n^\mu \rceil}$.

We introduce the L^1-norm

$$\|Z\|_1 := \mathbb{E}|Z|,$$

for each $\mathbb{R}^d$-valued random variable Z.

Lemma 3.3.4 *Under Assumptions 3.2.2 and 3.2.3, we have*

$$\sup_{t \geq 1} \|\overline{y}_t\|_1 + \sup_{t \geq 1} E\|\theta_t\|_1 < \infty.$$

Proof: Note that $2c_k H^j(\theta, x, c_k) = l_0(\theta + c_k \mathbf{e}_j) - l_0(\theta - c_k \mathbf{e}_j)$, for all x, $j = 1, \ldots, d$ when $\theta \notin B_{D+1}$. Furthermore, the function $l_0(\theta + c_k \mathbf{e}_j) - l_0(\theta - c_k \mathbf{e}_j)$ is Lipschitz on B_{D+1} which, together with Lemma 3.3.2 implies

$$\left\| \frac{l_0(\theta + c_k \mathbf{e}_j) - l_0(\theta - c_k \mathbf{e}_j)}{2c_k} - H^j(\theta, X_{k+1}, c_k) \right\|_1 \leq \bar{C}, \ \theta \in \mathbb{R}^d, \tag{3.8}$$

for a fixed constant $\bar{C}$. Clearly,

$$\begin{aligned} ||\theta_{k+1}||_1 \ &\leq \ \left\| \theta_k - \frac{\lambda_k}{2c_k} \sum_{j=1}^d [l_0(\theta_k + c_k \mathbf{e}_j) - l_0(\theta_k - c_k \mathbf{e}_j)] \mathbf{e}_j \right\|_1 \\ &+ \ \left\| \frac{\lambda_k}{2c_k} \sum_{j=1}^d [l_0(\theta_k + c_k \mathbf{e}_j) - l_0(\theta_k - c_k \mathbf{e}_j)] \mathbf{e}_j - \lambda_k H(\theta_k, X_{k+1}, c_k) \right\|_1 . \end{aligned}$$

Note that, by Assumption 3.2.2, l_0 is strongly convex, in particular,

$$\langle \nabla l_0(\theta) - \nabla l_0(0), \theta \rangle \geq A_0 |\theta|^2, \ \theta \in \mathbb{R}^d$$

for all θ, with some $A_0 > 0$. Hence also

$$\langle \nabla l_0(\theta), \theta \rangle \geq A|\theta|^2 - B, \ \theta \in \mathbb{R}^d$$

for suitable $A, B > 0$. But then for all $a > 0$ small enough,

$$|\theta - a\nabla l_0(\theta)| \leq (1 - A'a)|\theta| + aB', \ \theta \in \mathbb{R}^d$$

for suitable $A', B' > 0$. By the mean value theorem,

$$l_0(\theta_k + c_k\mathbf{e}_j) - l_0(\theta_k - c_k\mathbf{e}_j) = 2c_k\partial_j l_0(\xi_j)$$

for some random variable $\xi_j \in [\theta_k - c_k\mathbf{e}_j, \theta_k + c_k\mathbf{e}_j]$. Since ∇l_0 is Lipschitz,

$$\max_j \|\nabla l_0(\theta_k) - \nabla l_0(\xi_j)\|_1 \leq L'$$

for some $L' > 0$. It follows then easily that, for $k \geq k_0$ large enough such that λ_k is small enough,

$$\begin{aligned} & \left\| \theta_k - \frac{\lambda_k}{2c_k} \sum_{j=1}^{d} [l_0(\theta_k + c_k\mathbf{e}_j) - l_0(\theta_k - c_k\mathbf{e}_j)]\mathbf{e}_j \right\|_1 \\ \leq \ & \lambda_k dL' + ||\theta_k - \lambda_k \nabla l_0(\theta_k)||_1 \\ \leq \ & \lambda_k(B' + dL') + (1 - A'\lambda_k)||\theta_k||_1 \end{aligned}$$

holds. By (3.8),

$$\left\| \frac{\lambda_k}{2c_k} \sum_{j=1}^{d} [l_0(\theta_k + c_k\mathbf{e}_j) - l_0(\theta_k - c_k\mathbf{e}_j)]\mathbf{e}_j - \lambda_k H(\theta_k, X_{k+1}, c_k) \right\|_1 \leq \lambda_k d\bar{C}.$$

Apply Lemma 3.3.3 with the choice $x_k := ||\theta_k||_1$, $\underline{c} := d(L' + \bar{C}) + B'$ and $\zeta_k := \lambda_k$, $\nu := A'$ to obtain that $\sup_{k \geq k_0} \|\theta_k\|_1 < \infty$. Then trivially also $\sup_{n \in \mathbb{N}} \|\theta_n\|_1 < \infty$ holds, which easily implies $\sup_{t \geq 1} \|\theta_t\|_1 < \infty$ as well.

Now turning to $\overline{y}_t$ we see that, for $n \geq 1$ and $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$,

$$\begin{aligned} & |\overline{y}_t - \theta^*| = |\overline{y}_t - y(t, \lceil n^\mu \rceil, \theta^*)| \\ \leq \ & |\theta_{\lceil n^\mu \rceil} - \theta^*| C^*, \end{aligned}$$

finishing the proof. □

Lemma 3.3.5 *Let Assumptions 3.2.2 and 3.2.3 hold. Then there exists $C_l > 0$ such that $\sup_{k \geq 1} |J(\theta, X_k)| \leq C_l(1 + |\theta|^2)$, i.e. J grows at most quadratically in θ.*

Proof: Recall that

$$|J(\theta, x)| \leq |l_0(\theta)| + \sum_{i=1}^{m_s} 1_{A_i(x)} |l_i(x, \theta)|,$$

where the functions l_i are bounded on the bounded sets $\cup_{x \in \mathbb{R}^d} A_i(x)$ for $i = 1, \ldots, d$, and l_0 grows quadratically. □

The difficulty of the following lemma consists in handling the discontinuities and the dependence of the sequence X_k at the same time.

Lemma 3.3.6 *Let Assumptions 3.2.2 and 3.2.3 hold. Then for each $R > 0$ the random field $J(\theta, X_n)$, $\theta \in B_R$, $n \in \mathbb{N}$ satisfies*

$$M_3^n(J(\theta, X)) \leq C_l(1 + R^2),$$
$$\Gamma_3^n(J(\theta, X)) \leq L(1 + R^2),$$

for some $L > 0$, where C_l is as in Lemma 3.3.5.

Proof: The first statement follows from Lemma 3.3.5.

Let $n \geq 0$, $\tau \geq 1$ be fixed. For $k \geq \tau$, define $X_k^+ = \mathbb{E}[X_{n+k}|\mathcal{F}_{n+k-\tau}^+ \vee \mathcal{F}_n]$. For the sake of simplicity, assume that $m_s = 1$ in the definition of J, $m_p = 0$, but the same argument would work for several summands, too. We also take the process X unidimensional ($m := 1$) noting that the same arguments easily carry over to a general m.

We now perform an auxiliary estimate. Let $\epsilon_\tau > 0$ be a parameter to be chosen later and let $1 \leq j \leq m_p'$. We will write h below instead of h_1. Define $Z_k = X_{n+k} - X_k^+$ and estimate

$$\begin{aligned}
&\mathbb{E}_n \left| 1_{\{X_{n+k} > h^j(\theta)\}} - 1_{\{X_k^+ > h^j(\theta\})} \right|^3 = \mathbb{E}_n \left| 1_{\{X_{n+k} > h^j(\theta)\}} - 1_{\{X_k^+ > h^j(\theta\})} \right| \\
&\leq \mathbb{P}_n \left(X_{n+k} \in \left(h^j(\theta) - |Z_k|, h^j(\theta) + |Z_k| \right) \right) \\
&\leq \mathbb{P}_n \left(X_{n+k} \in \left(h^j(\theta) - |Z_k|, h^j(\theta) + |Z_k| \right), |Z_k| \leq \epsilon_\tau \right) + \mathbb{P}_n(|Z_k| \geq \epsilon_\tau) \\
&\leq 2F\epsilon_\tau + \frac{\mathbb{E}_n[|X_{n+k} - X_k^+|^3]}{\epsilon_\tau^3},
\end{aligned}$$

where the last inequality follows from Assumption 3.2.3 and the Markov inequality. Now estimate

$$\begin{aligned}
& \mathbb{E}_n^{1/3} \left| \left(\prod_{j=1}^{m_p'} 1_{\{X_{n+k} > h^j(\theta)\}} \right) l_1(X_{n+k}, \theta) - \left(\prod_{j=1}^{m_p'} 1_{\{X_k^+ > h^j(\theta)\}} \right) l_1(X_k^+, \theta) \right|^3 \\
\leq\ & \mathbb{E}_n^{1/3} \left| \left(\prod_{j=1}^{m_p'} 1_{\{X_{n+k} > h^j(\theta)\}} - \prod_{j=1}^{m_p'} 1_{\{X_k^+ > h^j(\theta)\}} \right) l_1(X_{n+k}, \theta) \right|^3 \\
+\ & \mathbb{E}_n^{1/3} \left| \left(l_1(X_{n+k}, \theta) - l_1(X_k^+, \theta) \right) \prod_{j=1}^{m_p'} 1_{\{X_k^+ > h^j(\theta)\}} \right|^3 \\
\leq\ & \mathbb{E}_n^{1/3} \left[\left| \sum_{j=1}^{m_p'} \left| 1_{\{X_{n+k} > h^j(\theta)\}} - 1_{\{X_k^+ > h^j(\theta)\}} \right| \right|^3 (L_1(|X_{n+k}| + R) + |l_1(0,0)|)^3 \right] \\
+\ & L_1 \mathbb{E}_n^{1/3} |X_{n+k} - X_k^+|^3 \\
\leq\ & C_1(1 + R) \left(\epsilon_\tau^{1/3} + \frac{\mathbb{E}_n^{1/3}[|X_{n+k} - X_k^+|^3]}{\epsilon_\tau} \right)
\end{aligned}$$

for some C_1, where we used the Lipschitz-continuity of the function l_1, as well as the observation that

$$\left|\prod_{j=1}^{m'_p} 1_{\{X_{n+k}>h^j(\theta)\}} - \prod_{j=1}^{m'_p} 1_{\{X_k^+>h^j(\theta)\}}\right| \leq \sum_{j=1}^{m'_p} \left|1_{\{X_{n+k}>h^j(\theta)\}} - 1_{\{X_k^+>h^j(\theta)\}}\right|.$$

A similar estimate works for l_0 but we get the upper bound

$$\begin{aligned} & \mathbb{E}_n^{1/3}\left|1_{A_0(X_{n+k})}l_0(\theta) - 1_{A_0(X_k^+)}l_0(\theta)\right|^3 \\ \leq\ & C_1(1+R^2)\left(\epsilon_\tau^{1/3} + \frac{\mathbb{E}_n^{1/3}[|X_{n+k}-X_k^+|^3]}{\epsilon_\tau}\right) \end{aligned}$$

instead. For the second inequality of the present lemma, note first that Lemma 3.4.1 below implies

$$\begin{aligned} & \mathbb{E}_n^{1/3}\left[\left|J(\theta,X_{n+k}) - \mathbb{E}[J(\theta,X_{n+k})|\mathcal{F}_n \vee \mathcal{F}^+_{n+k-\tau}]\right|^3\right] \\ \leq & 2\mathbb{E}_n^{1/3}\left[\left|J(\theta,X_{n+k}) - J(\theta,X_k^+)\right|^3\right], \end{aligned}$$

hence it suffices to estimate the latter quantity. From our previous estimates it follows that, for some $C>0$,

$$\mathbb{E}_n^{1/3}\left[|J(\theta,X_{n+k}) - J(\theta,X_k^+)|^3\right] \leq C(1+R^2)\left[\sqrt[3]{\epsilon_\tau} + \frac{\mathbb{E}_n^{1/3}|X_{n+k}-X_k^+|^3}{\epsilon_\tau}\right] \tag{3.9}$$

Choose $\epsilon_\tau := (\tau+1)^{-3-\epsilon/2}$. Summing up the right-hand side for $\tau \geq 1$ we see that, by Assumption 3.2.4, the sum has an upper bound independent of k. The statement follows as the case $\tau = 0$ is easy. □

3.3.2 Decreasing gain case

The following lemma contains the core estimates of the present paper.

Lemma 3.3.7 *Let $n \geq 1$. Let $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$ for $\mu := 1/\gamma$ and let Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5 and 3.2.6 hold. Then $\mathbb{E}|\theta_t - \overline{y}_t| = O(n^{-\beta})$, where $\beta = \min\left(\frac{1}{2\gamma} - \frac{1}{2}, 2\right)$.*

Proof: For $\lceil n^\mu \rceil \le t < \lceil (n+1)^\mu \rceil$,

$$\begin{aligned}
|\theta_{\lfloor t \rfloor} - \overline{y}_t| &\le |\overline{y}_{\lfloor t \rfloor} - \overline{y}_t| + |\theta_{\lfloor t \rfloor} - \overline{y}_{\lfloor t \rfloor}| \\
&\le \int_{\lfloor t \rfloor}^{t} a_u \left| G(\overline{y}_u) \right| \, du + \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H(u, \theta_{\lfloor u \rfloor}) - G(\overline{y}_u) \right) \, du \right| \\
&\le a_{n^\mu} \int_{\lfloor t \rfloor}^{t} \left| G(\overline{y}_u) \right| \, du \\
&+ \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H(u, \theta_{\lfloor u \rfloor}) - H(u, \overline{y}_u) \right) \, du \right| \\
&+ \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H(u, \overline{y}_u) - \mathbb{E}[H(u, \overline{y}_u) | \mathcal{F}_{\lceil n^\mu \rceil}] \right) \, du \right| \\
&+ \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(\mathbb{E}[H(u, \overline{y}_u) | \mathcal{F}_{\lceil n^\mu \rceil}] - G(\overline{y}_u) \right) \, du \right| \\
&=: \Sigma_0 + \Sigma_1 + \Sigma_2 + \Sigma_3.
\end{aligned}$$

Estimation of Σ_0. Since G has at most linear growth, Lemma 3.3.4 guarantees that

$$\mathbb{E}[\Sigma_0] = O\left(a_{n^\mu} \int_{\lfloor t \rfloor}^{t} (\mathbb{E}|\overline{y}_u| + 1) \, du \right) = O(n^{-\mu}).$$

Estimation of Σ_1. Recall that, by the tower property for conditional expectations,

$$\mathbb{E} \left| H(u, \theta_u) - H(u, \overline{y}_u) \right| = \mathbb{E}\mathbb{E}_k \left| H(u, \theta_u) - H(u, \overline{y}_u) \right|$$

for all integers k. Applying this observation to $k = \lfloor u \rfloor$, Lemma 3.3.1 implies that

$$\begin{aligned}
\mathbb{E}[\Sigma_1] &= \mathbb{E} \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H(u, \theta_{\lfloor u \rfloor}) - H(u, \overline{y}_u) \right) \, du \right| \\
&\le \int_{\lceil n^\mu \rceil}^{t} a_u \mathbb{E} \left| H(u, \theta_{\lfloor u \rfloor}) - H(u, \overline{y}_u) \right| \, du \\
&\le C_\flat \int_{\lceil n^\mu \rceil}^{t} \frac{a_u}{c_u} \mathbb{E} \left| \theta_{\lfloor u \rfloor} - \overline{y}_u \right| \, du + C_\flat \int_{\lceil n^\mu \rceil}^{t} \frac{a_u}{c_u} c_u^2 \, du
\end{aligned} \tag{3.10}$$

Henceforth we will denote

$$\Sigma_1' := C_\flat \int_{\lceil n^\mu \rceil}^{t} \frac{a_u}{c_u} c_u^2 \, du.$$

Notice that

$$E[\Sigma_1'] = O(n^{-\mu\gamma-1}) = O(n^{-2}).$$

Estimation of Σ_2. Notice that $H(u,\bar{\theta}) = \mathbb{E}[H(u,\bar{\theta})|\mathcal{F}_{\lceil n^\mu \rceil}]$ for all $\mathcal{F}_{\lceil n^\mu \rceil}$-measurable $\bar{\theta}$ such that a.s. $\bar{\theta} \notin B_D$ since $J(\theta,x)$ does not depend on x outside B_D by Assumption 3.2.2. Thus

$$\Sigma_2 \leq \sup_{\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil} \left| \int_{\lceil n^\mu \rceil}^{t} a_u \mathbb{1}_{\{\overline{y}_u \in B_D\}} \left(H(u,\overline{y}_u) - \mathbb{E}[H(u,\overline{y}_u)|\mathcal{F}_{n^\mu}] \right) du \right|.$$

We will use the inequality of Theorem 3.4.1 below with $r=3$, with $\mathcal{R}_t := \mathcal{F}_{t+\lceil n^\mu \rceil}$, $t \in \mathbb{R}_+$, $\mathcal{R}_t^+ := \mathcal{F}^+_{t+\lceil n^\mu \rceil}$ with the process defined by

$$W_t = \mathbb{1}_{\{\overline{y}_{t+\lceil n^\mu \rceil} \in B_D\}} c_{t+\lceil n^\mu \rceil} \left(H(t,\overline{y}_{t+\lceil n^\mu \rceil}) - \mathbb{E}[H(t,\overline{y}_{t+\lceil n^\mu \rceil})|\mathcal{F}_{\lceil n^\mu \rceil}] \right), \ t \geq 0 \tag{3.11}$$

and with the function $f_t = a_{t+\lceil n^\mu \rceil}/c_{t+\lceil n^\mu \rceil}$. Note that $\{\overline{y}_t \in B_D\} \in \mathcal{F}_{\lceil n^\mu \rceil}$ for all $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$. We get from Lemma 3.4.2 below and from the cited inequality that

$$\begin{aligned}
\mathbb{E}[\Sigma_2] &= \mathbb{E}[\mathbb{E}[\Sigma_2|\mathcal{F}_{\lceil n^\mu \rceil}]] \leq \mathbb{E}[\mathbb{E}^{1/3}[\Sigma_2^3|\mathcal{F}_{\lceil n^\mu \rceil}]] \\
&\leq C'(3) \left(\int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \left(\frac{a_u}{c_u} \right)^2 du \right)^{1/2} \mathbb{E}[\tilde{M}_3 + \tilde{\Gamma}_3] \\
&\leq C'(3) \left(\int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \left(\frac{a_u}{c_u} \right)^2 du \right)^{1/2} C(1+D^2).
\end{aligned}$$

We thus get

$$\mathbb{E}[\Sigma_2] = O\left(\int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \left(\frac{a_u}{c_u} \right)^2 du \right)^{1/2}.$$

Estimation of Σ_3.

$$\begin{aligned}
\mathbb{E}[\Sigma_3] &\le \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^{t} a_u \left|\mathbb{E}[H(u,\overline{y}_u)|\mathcal{F}_{\lceil n^\mu\rceil}] - G(\overline{y}_u)\right| du\right] \\
&\le \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^{t} a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}[H(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor})|\mathcal{F}_{\lceil n^\mu\rceil}] - G(\vartheta)\right| du\right] \\
&\le \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^{t} a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}[H(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor})|\mathcal{F}_{\lceil n^\mu\rceil}] - \mathbb{E}[H(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor})]\right| du\right] \\
&+ \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^{t} a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}[H(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor})] - G(\vartheta)\right| du\right] \qquad (3.12)
\end{aligned}$$

To handle the second sum, note that, for each $i = 1, \dots, d$,

$$\mathbb{E}[H^i(\vartheta, X_{k+1}, c_k)] = \frac{U(\vartheta + c_k\mathbf{e}_i) - U(\vartheta - c_k\mathbf{e}_i)}{2c_k} = G^i(\xi_k^i),$$

for some $\xi_k^i \in [\vartheta - c_k\mathbf{e}_i, \vartheta + c_k\mathbf{e}_i]$. The Lipschitz continuity of G implies that $|G^i(\xi_k^i) - G^i(\vartheta)| \le L_G c_k$ so

$$\begin{aligned}
&\mathbb{E}\left[\int_{\lceil n^\mu\rceil}^{t} a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}[H(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor})] - G(\vartheta)\right| du\right] \le \int_{\lceil n^\mu\rceil}^{t} a_u d L_G c_{\lceil u\rceil}\, du \\
&= O\left(\int_{\lceil n^\mu\rceil}^{\lceil (n+1)^\mu\rceil} u^{-1-\gamma}\, du\right) = O(n^{-2}).
\end{aligned}$$

Now we turn to the first sum in (3.12). Define $X_k^+ = \mathbb{E}[X_k|\mathcal{F}^+_{\lceil n^\mu\rceil}]$, $k \ge \lceil n^\mu\rceil$. First let us estimate

$$\mathbb{E}_{\lceil n^\mu\rceil}[|H(\vartheta, X_{k+1+\lceil n^\mu\rceil}, c_{k+\lceil n^\mu\rceil}) - H(\vartheta, X^+_{k+1+\lceil n^\mu\rceil}, c_{k+\lceil n^\mu\rceil})|].$$

Fix $\epsilon_k > 0$ to be chosen later. By an argument similar to that of Lemma 3.3.6 (using the first instead of the third moment in Markov's inequality) we get that, for some constant C_1,

$$\begin{aligned}
&c_{k+\lceil n^\mu\rceil}\mathbb{E}_{\lceil n^\mu\rceil}[|H(\vartheta, X_{\lceil n^\mu\rceil+k+1}, c_{k+\lceil n^\mu\rceil}) - H(\vartheta, X^+_{\lceil n^\mu\rceil+k+1}, c_{k+\lceil n^\mu\rceil})|] \\
\le\ & C_1\left[\epsilon_k + \frac{\mathbb{E}_{\lceil n^\mu\rceil}[|X_{\lceil n^\mu\rceil+k+1} - X^+_{\lceil n^\mu\rceil+k+1}|]}{\epsilon_k}\right].
\end{aligned}$$

Choose $\epsilon_k = (1+k)^{-1-\epsilon/2}$. Then using Assumption 3.2.4 we get

$$c_{k+\lceil n^\mu \rceil} \sup_{\vartheta \in \mathbb{R}^d} \mathbb{E}[|H(\vartheta, X_{\lceil n^\mu \rceil + k+1}, c_{\lceil n^\mu \rceil + k}) - H(\vartheta, X^+_{\lceil n^\mu \rceil + k+1}, c_{k+\lceil n^\mu \rceil})||\mathcal{F}_{\lceil n^\mu \rceil}] = O(k^{-1-\epsilon/2})$$

which also implies

$$c_{k+\lceil n^\mu \rceil} \sup_{\vartheta \in \mathbb{R}^d} \mathbb{E}[|H(\vartheta, X_{\lceil n^\mu \rceil + k+1}, c_{\lceil n^\mu \rceil + k}) - H(\vartheta, X^+_{\lceil n^\mu \rceil + k+1}, c_{\lceil n^\mu \rceil + k})|] = O(k^{-1-\epsilon/2}).$$

Since $\mathbb{E}[H(\vartheta, X^+_{k+1}, c_{\lceil n^\mu \rceil + k})|\mathcal{F}_{\lceil n^\mu \rceil}] = \mathbb{E}[H(\vartheta, X^+_{k+1}, c_{k+\lceil n^\mu \rceil})]$ for $k \geq \lceil n^\mu \rceil$ by independence of $\mathcal{F}_{\lceil n^\mu \rceil}$ and $\mathcal{F}^+_{\lceil n^\mu \rceil}$, we have

$$\begin{aligned}
&\int_{\lceil n^\mu \rceil}^{t} a_u \mathbb{E} \sup_{\vartheta \in \mathbb{R}^d} \left| \mathbb{E}[H(\vartheta, X_{\lfloor u \rfloor + 1}, c_{\lfloor u \rfloor})|\mathcal{F}_{\lceil n^\mu \rceil}] - \mathbb{E}[H(\vartheta, X_{\lfloor u \rfloor + 1}, c_{\lfloor u \rfloor})] \right| du \\
&\leq \int_{\lceil n^\mu \rceil}^{\infty} \frac{a_u}{c_u} c_u \mathbb{E} \sup_{\vartheta \in \mathbb{R}^d} \mathbb{E}[\left| H(\vartheta, X_{\lfloor u \rfloor + 1}, c_k) - H(\vartheta, X^+_{\lfloor u \rfloor + 1}, c_k) \right| |\mathcal{F}_{\lceil n^\mu \rceil}] \, du \\
&+ \int_{\lceil n^\mu \rceil}^{\infty} \frac{a_u}{c_u} c_u \sup_{\vartheta \in \mathbb{R}^d} \mathbb{E}\left[\left| H(\vartheta, X_{\lfloor u \rfloor + 1}, c_k) - H(\vartheta, X^+_{\lfloor u \rfloor + 1}, c_k) \right| \right] du \\
&\leq C_2 \frac{a_{\lceil n^\mu \rceil}}{c_{\lceil n^\mu \rceil}} \sum_{k=1}^{\infty} k^{-1-\epsilon/2}
\end{aligned}$$

with some C_2, so

$$\mathbb{E}[\Sigma_3] = O\left(n^{\mu(\gamma-1)}\right).$$

Combining the estimates we have so far, we get

$$E[\Sigma_0 + \Sigma_1' + \Sigma_2 + \Sigma_3] = O(n^{-\mu} + n^{-2} + n^{\frac{-\mu+2\mu\gamma-1}{2}} + n^{-2} + n^{\mu(\gamma-1)}). \tag{3.13}$$

Notice that $\mathbb{E}|\theta_t - \overline{y}_t|$ is always finite, see Lemma 3.3.4 above. Use Gronwall's lemma and (3.10) to obtain the inequality

$$\mathbb{E}[|\theta_{\lfloor t \rfloor} - \overline{y}_t|] \leq E[\Sigma_0 + \Sigma_1' + \Sigma_2 + \Sigma_3] \exp\left(C_3 \int_{n^\mu}^{\lceil (n+1)^\mu \rceil} \frac{a_u}{c_u} \, du \right)$$

with some constant C_3. From Lemma 3.3.2 it is also easy to check that $\mathbb{E}|\theta_t - \theta_{\lfloor t \rfloor}| = O(n^{-\mu})$. Note furthermore that the terms $n^{-\mu}$ and $n^{\mu(\gamma-1)}$ are always negligible in (3.13). These observations lead to

$$\begin{aligned}
\mathbb{E}|\theta_t - \overline{y}_t| &= O(n^{\frac{-\mu+2\mu\gamma-1}{2}} + n^{-2}) \exp\left(C_4 n^{\mu\gamma-1}\right) \\
&= O(n^{\frac{1}{2}-\frac{1}{2\gamma}} + n^{-\mu\gamma-1})
\end{aligned}$$

with some C_4, finsihing the proof. □

Proof: [Proof of Theorem 3.2.1] Denote

$$d_i = \sup_{\lceil i^\mu \rceil \le s < \lceil (i+1)^\mu \rceil} \mathbb{E}|\theta_s - \overline{y}_s|, \ i = 1, 2, \dots$$

By Fatou's lemma, we also have

$$\mathbb{E}|\theta_{\lceil (i+1)^\mu \rceil} - \overline{y}_{\lceil (i+1)^\mu \rceil -}| \le d_i$$

where $\overline{y}_{s-}$ denotes the left limit of $\overline{y}$ at s.

It follows from Lemma 3.3.7, that $d_i = O(i^{-\beta})$. Combining this with Assumption 3.2.6 and using telescoping sums we get, for each integer $N \ge 1$,

$$\begin{aligned}
&\mathbb{E}|y(\lceil N^\mu \rceil, 1, \theta_1) - \theta_{\lceil N^\mu \rceil}| \\
&= \mathbb{E}|y(\lceil N^\mu \rceil, 1, \theta_1) - y(\lceil N^\mu \rceil, \lceil N^\mu \rceil, \theta_{\lceil N^\mu \rceil})| \\
&\le \sum_{i=2}^{N} \mathbb{E}\left|y(\lceil N^\mu \rceil, \lceil (i-1)^\mu \rceil, \theta_{\lceil (i-1)^\mu \rceil}) - y(\lceil N^\mu \rceil, \lceil i^\mu \rceil, \theta_{\lceil i^\mu \rceil})\right| \\
&\le \sum_{i=2}^{N} \mathbb{E}\left|y(\lceil N^\mu \rceil, \lceil i^\mu \rceil, y(\lceil i^\mu \rceil, \lceil (i-1)^\mu \rceil, \theta_{\lceil (i-1)^\mu \rceil})) - y(\lceil N^\mu \rceil, \lceil i^\mu \rceil, \theta_{\lceil i^\mu \rceil})\right| \\
&\le C^* \sum_{i=2}^{N} \left(\frac{i+1}{N}\right)^{\alpha\mu} d_{i-1} = O(N^{-\beta+1}),
\end{aligned}$$

noting that $y(\lceil i^\mu \rceil, \lceil (i-1)^\mu \rceil, \theta_{\lceil (i-1)^\mu \rceil})$ equals the left limit $\overline{y}_{\lceil i^\mu \rceil -}$. A similar argument provides, for all $t \in (\lceil N^\mu \rceil, \lceil (N+1)^\mu \rceil)$,

$$\mathbb{E}|\theta_t - y(t, 1, \theta_1)| = O(N^{-\beta+1}).$$

Taking μth root we obtain

$$\mathbb{E}|\theta_t - y(t, 1, \theta_1)| = O(t^{\frac{-\beta+1}{\mu}}), \ t \ge 1.$$

To conclude, note that by the stability Assumption 3.2.6, $|y(t, 1, \theta_1) - \theta^*| \le C^*|\theta_1 - \theta^*|t^{-\alpha}$ and that $E|\theta_1| < \infty$, as easily seen using Lemma 3.3.2. □

3.3.3 Fixed gain case

Define $T = \frac{c}{a}$. For $nT \le t < (n+1)T$, define $\overline{y}_t = y(t, nT, \theta_{nT})$, i.e. the solution of (3.5) with the initial condition $y_{nT} = \theta_{nT}$. We use the piece-wise linear extension $\overline{\theta}_t$ of θ_t and the piece-wise constant extension $H(t, \theta)$ of $H(\theta, X_{k+1}, c)$ as defined in the decreasing gain setting, but a and c are now constants.

Lemma 3.3.8 *Let Assumptions 3.2.1,3.2.2, 3.2.3 3.2.4 and 3.2.6 hold. Then for $t \in [nT, (n+1)T]$ there is $\overline{C} > 0$ such that $\mathbb{E}|\theta_t - \overline{y}_t| \le \overline{C} \max(c^2, \sqrt{\frac{a}{c}})$.*

Proof: Using essentially the same estimates we derived in the decreasing gain setting, for fixed a and c we get

$$\mathbb{E}[\Sigma_0] \leq C_0 a \tag{3.14}$$

$$\mathbb{E}[\Sigma_1] \leq C_1 \left[\frac{a}{c} \sum_{nT}^{t-1} \mathbb{E}|\theta_k - \overline{y_k}| + c^2 \right] \tag{3.15}$$

$$\mathbb{E}[\Sigma_2] \leq C_2 \left(\sum_{nT}^{t-1} \left(\frac{a^2}{c^2} \right) \right)^{1/2} \leq C_2 \left(\frac{c}{a} \frac{a^2}{c^2} \right)^{1/2} = C_2 \sqrt{\frac{a}{c}} \tag{3.16}$$

$$\mathbb{E}[\Sigma_3] \leq C_3 \left[\frac{a}{c} + \sum_{nT}^{t-1} ac \right] = C_3 T a c + C_3 \frac{a}{c} = O\left(c^2 + \frac{a}{c} \right), \tag{3.17}$$

with suitable constants C_0, C_1, C_2, C_3. Combine these estimates and use Gronwall-lemma to get the statement. To choose optimally, set $c^2 = \sqrt{\frac{a}{c}}$, that is $c = a^{\frac{1}{5}}$. In this case $\mathbb{E}|\theta_t - \overline{y}_t| \leq C_4 a^{\frac{2}{5}}$ for some C_4. □

Proof: [Proof of Theorem 3.2.2] Denote

$$d_i = \sup_{iT \leq s < (i+1)T} \mathbb{E}|\theta_s - \overline{y}_s^i|.$$

It follows from Lemma 3.3.8, that $d_i \leq \overline{C} \max(c^2, \sqrt{\frac{a}{c}})$. Combining this with Assumption 3.2.6 and using telescoping sums we get

$$\begin{aligned}
\mathbb{E}|y(NT, 1, \theta_1) - \theta_{NT}| &= \mathbb{E}|y(NT, 1, \theta_1) - y(NT, NT, \theta_{NT})| \\
&\leq \sum_{i=2}^{N} \mathbb{E}\left|y(NT, (i-1)T, \theta_{(i-1)T}) - y(NT, iT, \theta_{iT})\right| \\
&\leq \sum_{i=2}^{N} \mathbb{E}\left|y(NT, iT, y(iT, (i-1)T, \theta_{(i-1)T})) - y(NT, iT, \theta_{iT})\right| \\
&\leq \sum_{i=2}^{N} \left(C^* e^{-a\alpha(NT-iT)}\right) d_{i-1} \leq \hat{C} \max\left(c^2, \sqrt{\frac{a}{c}}\right),
\end{aligned}$$

with some $\hat{C}$ since $\sum_{i=2}^{N} e^{-a\alpha(NT-iT)}$ has an upper bound independent of N. We similarly get

$$\sup_{NT \leq t < (N+1)T} \mathbb{E}|\theta_t - y(t, 1, \theta_1)| \leq \check{C} \max\left(c^2, \sqrt{\frac{a}{c}}\right)$$

with some $\check{C}$. To conclude, note that by the stability Assumption 3.2.6, $|y(t, 1, \theta_1) - \theta^*| \leq C^* |\theta_1 - \theta^*| e^{-a\alpha t}$ and therefore

$$\mathbb{E}|\theta_t - \theta^*| = O\left(\max\left(c^2, \sqrt{\frac{a}{c}}\right) + e^{-a\alpha t} \right).$$

□

3.4 Auxiliary results

We define continuous-time analogues of the key quantities M and Γ from Assumption 3.2.4 and establish a pivotal maximal inequality for them.

Consider a continuous-time filtration $(\mathcal{R}_t)_{t\in\mathbb{R}_+}$ as well as a decreasing family of sigma-fields $(\mathcal{R}_t^+)_{t\in\mathbb{R}_+}$. We assume that $\mathcal{R}_t$ is independent of $\mathcal{R}_t^+$, for all $t \in \mathbb{R}_+$.

We consider an $\mathbb{R}^d$-valued continuous-time stochastic process $(W_t)_{t\in\mathbb{R}_+}$ which is progressively measurable (i.e. $W : [0,t] \times \Omega \to \mathbb{R}^d$ is $\mathcal{B}([0,t]) \otimes \mathcal{R}_t$-measurable for all $t \in \mathbb{R}_+$).

From now on we assume that $W_t \in L^1$, $t \in \mathbb{R}_+$. Fix $r \geq 1$. We define the quantities

$$\begin{aligned} \tilde{M}_r &:= \operatorname*{ess\,sup}_{t\in\mathbb{R}_+} \mathbb{E}^{1/r}\left[|W_t|^r|\mathcal{R}_0\right], \\ \tilde{\gamma}_r(\tau) &:= \operatorname*{ess\,sup}_{t\geq\tau} \mathbb{E}^{1/r}[|W_t - \mathbb{E}[W_t|\mathcal{R}_{t-\tau}^+ \vee \mathcal{R}_0]|^r|\mathcal{R}_0], \ \tau \in \mathbb{R}_+, \end{aligned}$$

and set $\tilde{\Gamma}_r := \sum_{\tau=0}^{\infty} \tilde{\gamma}_r(\tau)$.

Now we recall a powerful maximal inequality, Theorem B.3 of [Barkhagen et al., 2019].

Theorem 3.4.1 *Let $(W_t)_{t\in\mathbb{R}_+}$ be L^r-bounded for some $r > 2$ and let $\tilde{M}_r + \tilde{\Gamma}_r < \infty$ a.s. Assume $\mathbb{E}[W_t|\mathcal{R}_0] = 0$ a.s. for $t \in \mathbb{R}_+$. Let $f : [0,T] \to \mathbb{R}$ be $\mathcal{B}([0,T])$-measurable with $\int_0^T f_t^2\,dt < \infty$. Then there is a constant $C'(r)$ such that*

$$\mathbb{E}^{1/r}\left[\sup_{s\in[0,T]}\left|\int_0^s f_t W_t\,dt\right|^r|\mathcal{R}_0\right] \leq C'(r)\left(\int_0^T f_t^2\,dt\right)^{1/2}[\tilde{M}_r + \tilde{\Gamma}_r], \tag{3.18}$$

almost surely. □

We also recall Lemma A.1 of [Chau et al., 2019a].

Lemma 3.4.1 *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be sigma-algebras. Let $X, Y \in \mathbb{R}^d$ be random variables in L^p such that Y is measurable with respect to $\mathcal{H} \vee \mathcal{G}$. Then for any $p \geq 1$,*

$$E^{1/p}\left[|X - E[X|\mathcal{H} \vee \mathcal{G}]|^p\middle|\mathcal{G}\right] \leq 2E^{1/p}\left[|X - Y|^p\middle|\mathcal{G}\right].$$

□

Lemma 3.4.2 *Let the process W_t be defined by (3.11). Taking the filtration $\mathcal{R}_t := \mathcal{F}_{t+\lceil n^\mu\rceil}$ and $\mathcal{R}_t^+ := \mathcal{F}^+_{t+\lceil n^\mu\rceil}$, we get $\tilde{M}(W) + \tilde{\Gamma}(W) \leq C(1 + D^2)$ for some $C > 0$.*

Proof: Note that the multiplication with the indicator function can only reduce the values, so we leave that away. We will prove this statement in the one-dimensional case, the several dimensional follows similarly. By the definition of H we can write

$$\begin{aligned}
&c_{t+\lceil n^\mu \rceil} \left(H(t, \overline{y}_{t+\lceil n^\mu \rceil}) - \mathbb{E}[H(t, \overline{y}_{t+\lceil n^\mu \rceil}) | \mathcal{F}_{\lceil n^\mu \rceil}] \right) \\
&= \frac{1}{2}(J(\theta + c_{t+\lceil n^\mu \rceil}, \overline{y}_{t+\lceil n^\mu \rceil}) - \mathbb{E}[J(\theta + c_{t+\lceil n^\mu \rceil}, \overline{y}_{t+\lceil n^\mu \rceil}) | \mathcal{F}_{\lceil n^\mu \rceil}]) \\
&- \frac{1}{2}(J(\theta - c_{t+\lceil n^\mu \rceil}, \overline{y}_{t+\lceil n^\mu \rceil}) - \mathbb{E}[J(\theta - c_{t+\lceil n^\mu \rceil}, \overline{y}_{t+\lceil n^\mu \rceil}) | \mathcal{F}_{\lceil n^\mu \rceil}])
\end{aligned}$$

,

Denote $W_t^{\pm} = J(\theta \pm c_{t+\lceil n^\mu \rceil}, \overline{y}_{t+\lceil n^\mu \rceil}) - \mathbb{E}[J(\theta \pm c_{t+\lceil n^\mu \rceil}, \overline{y}_{t+\lceil n^\mu \rceil}) | \mathcal{F}_{\lceil n^\mu \rceil}]$ and set $R = D$. Lemma 3.3.1, Lemma A.3 and Remark A.4 of [Chau et al., 2019a] imply that $\tilde{M}(W^{\pm}) \leq 2\tilde{M}(J)$ and $\tilde{\Gamma}(W^{\pm}) = \tilde{\Gamma}(J)$. Then use the estimates of Lemma 3.3.6 to conclude. □

3.5 Numerical experiments

In what follows we present numerical results to check the convergence of the algorithm for a simple discontinuous function J, defined as

$$J(\theta, X) = \begin{cases} (\theta - X)^2 + 1, \text{ if } X \leq \theta \\ (\theta - X)^2, \text{ otherwise,} \end{cases}$$

where X is a square-integrable, absolutely continuous random variable. Clearly, this function is not continuous in the parameter, but its expectation *is* continuous:

$$\begin{aligned}
U(\theta) = \mathbb{E}J(X, \theta) &= \int_{-\infty}^{\theta} ((x - \theta^2) + 1) f(x) dx + \int_{\theta}^{\infty} (x - \theta)^2 f(x) dx \\
&= \mathbb{E}(X - \theta)^2 + F(\theta) = \mathbb{E}X^2 - 2\theta \mathbb{E}X + \theta^2 + F(\theta),
\end{aligned}$$

where $f(\cdot)$ and $F(\cdot)$ are the density function resp. the distribution function of X. See Figure 3.1 below to see a visualization of such function. Assuming that F is differentiable, we need to solve

$$\frac{\partial U(\theta)}{\partial \theta} = -2\mathbb{E}X + 2\theta - f(\theta) = 0$$

in order to find $\arg\min \mathbb{E}J(X, \theta)$.

For the numerical examples we will use the recursion

$$\theta_{k+1} = \theta_k + \frac{1}{k + k_0} \frac{J(\theta + (k + k_0)^{-1/5}, X_{k+1}) - J(\theta - (k + k_0)^{-1/5}, X'_{k+1})}{(k + k_0)^{-1/5}}. \tag{3.19}$$

To compute the expected error, Monte Carlo simulations were used with 10000 sample paths and the number of steps k ranging from 2^8 to 2^{20}. We fit regression on the log-log plot to get the convergence rate only on $[2^{13}, 2^{20}]$ and set $k_0 = 10000$ to avoid the initial fluctuations of the algorithm. More technical remarks on the numerical examples are included in A

3.5.1 Independent innovations

In this section we assume that the consecutive "measurement noises" X_n are i.i.d. We consider three different choices for the distribution of the noise: normal, uniform and beta distributions. Note that normal distribution violates boundedness and for uniform distribution the differentiability of F fails, however convergence is achieved even in these cases. We also distinguish between the case where the observations X_{k+1} and X'_{k+1} are the same and when they are independent. Here we refer back to Remark 3.2.1 where we point out that this choice does not influence our theoretical results, however it may make a visible difference numerically. This phenomenon has already been observed, see [Glasserman and Yao, 1992,Spall, 2005] for more about the variance reduction technique called *common random numbers* (CRN). The values in Table 3.1 below represent the slope of linear regression we fit on the log-log plot of the average absolute error vs. the number of steps, together with the R-squared value measuring the goodness of the fit.

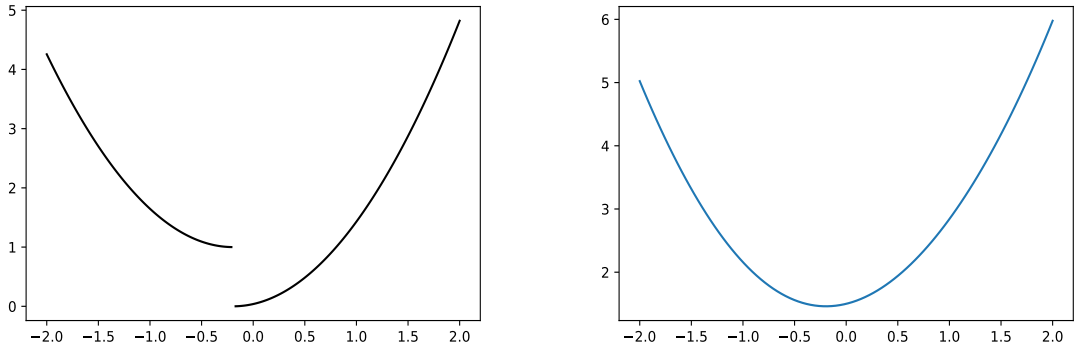
	independent X_{k+1},X'_{k+1}	identical $X_{k+1} = X'_{k+1}$
N$(0,1)$	-0.299 $(R^2 = 0.999)$	-0.459 $(R^2 = 0.999)$
U$([0,1])$	-0.14 $(R^2 = 0.997)$	-0.14 $(R^2 = 0.997)$
Beta$(2,2)$	-0.374 $(R^2 = 0.999)$	-0.393 $(R^2 = 0.999)$

Table 3.1: Convergence speed for different distributions of i.i.d. noise

The lower limit that we theoretically achieved for the convergence rate in Theorem 3.2.1 was -0.2, however the numerical experiments we present show that the practical convergence rate can outperform this.

Standard normal distribution

Assume that $X \sim N(0,1)$. Then the function we aim to find the minimum of is $U_1(\theta) = 1 + \theta^2 + \Phi(\theta)$, where Φ denotes the the cumulative distribution function of standard normal distribution. We get the solution $\theta^* = -\sqrt{W\left(\frac{1}{8\pi}\right)} \approx -0.19569$, where W is the Lambert-W function.

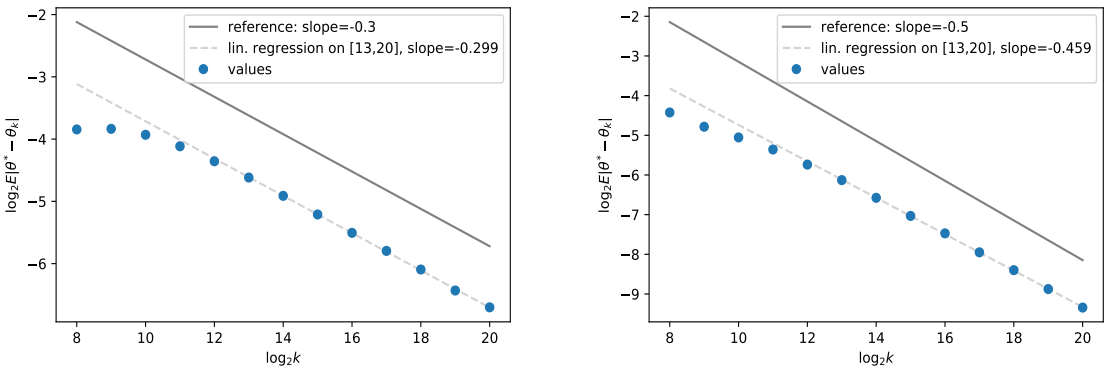


(a) $J(\theta^*, X)$ for the optimal θ^*

(b) $U_1(\theta) = \mathbb{E}[J(\theta, X)]$

Figure 3.1: The discontinuous stochastic representation and the smooth objective function

Figure 3.2 illustrates the convergence of two variations of algorithm (3.19) for U_1, starting the iteration from $\theta_0 = -0.1$. On figure (3.2a) we present the case where X_{k+1} and X'_{k+1} are independent on a log-log plot, we observe a convergence rate of $k^{-0.299}$ while (3.2b) shows the case where $X_{k+1} = X'_{k+1}$ which yields a convergence rate of $k^{-0.459}$.



(a) X_{k+1} and X'_{k+1} independent

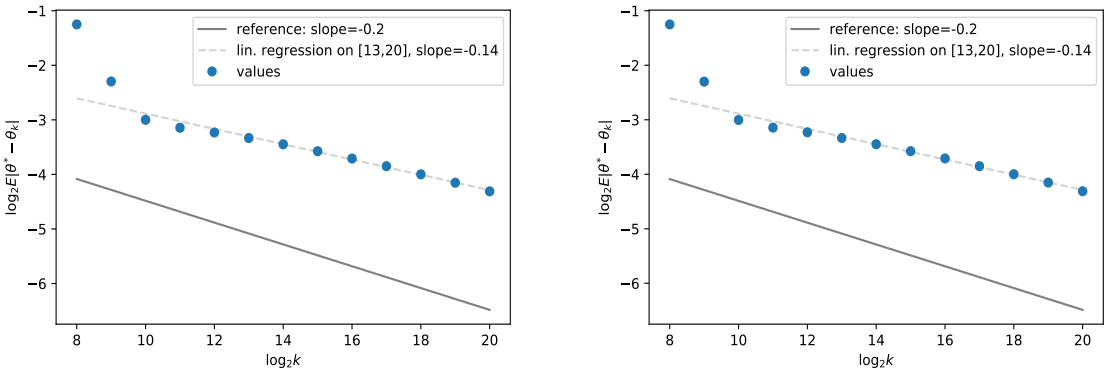
(b) $X_{k+1} = X'_{k+1}$

Figure 3.2: Log-log plot of $\mathbb{E}|\theta^* - \theta_k|$ vs. number of iterations for i.i.d. standard normal innovations

Uniform([0,1]) distribution

Let $X \sim \text{Uniform}([0, 1])$. Then the function we aim to find the minimum of is $U_2(\theta) = 1/3 - \theta + \theta^2 + F_{\text{uni}}(\theta)$, where F_{uni} denotes the the cumulative distribution function of Uniform([0,1]) distribution. We get the solution $\theta^* = 0$.

Figure 3.3 illustrates the convergence of two variations of algorithm (3.19) for U_2, starting the iteration from $\theta_0 = 1$. On figure (3.3a) we present the case where X_{k+1} and X'_{k+1} are independent on a log-log plot, while (3.3b) shows the case where $X_{k+1} = X'_{k+1}$, both of which yield a convergence rate of $k^{-0.14}$, worse that the theoretical rate $k^{-0.2}$.



(a) X_{k+1} and X'_{k+1} independent

(b) $X_{k+1} = X'_{k+1}$

Figure 3.3: Log-log plot of $\mathbb{E}|\theta^* - \theta_k|$ vs. number of iterations for i.i.d. uniform innovations

Beta(2,2) distribution

Let $X \sim \text{Beta}(2, 2)$. Then the function we aim to find the minimum of is $U_3(\theta) = 0.3 - \theta + \theta^2 + F_\beta(\theta)$, where F_β denotes the the cumulative distribution function of Beta(2,2) distribution. We get the solution $\theta^* = \frac{2-\sqrt{2.5}}{3} \approx 0.13962$.

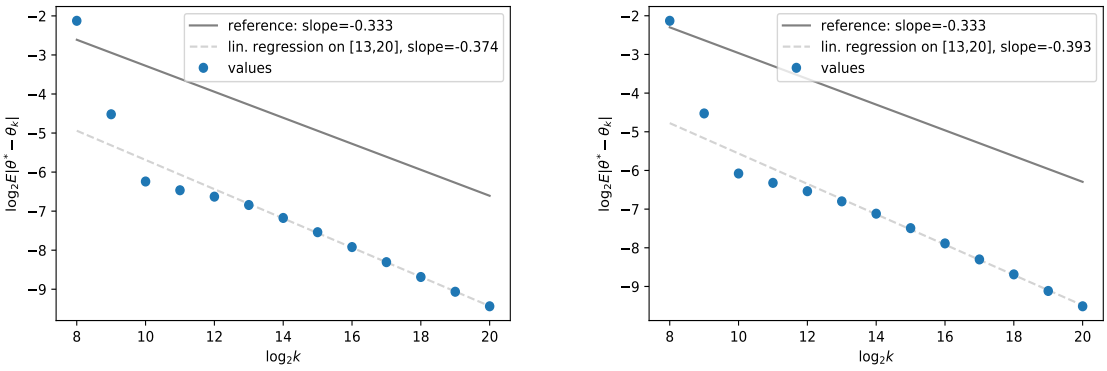
Figure 3.4 illustrates the convergence of two variations of algorithm (3.19) for U_3, starting the iteration from $\theta_0 = 1$. On figure (3.4a) we present the case where X_{k+1} and X'_{k+1} are independent on a log-log plot, we observe a convergence rate of $k^{-0.374}$ while (3.4b) shows the case where $X_{k+1} = X'_{k+1}$ which yields a convergence rate of $k^{-0.393}$.

3.5.2 AR(1) innovations

For an example with non-i.i.d. X_t, assume that the "noise" is an AR(1) process defined as

$$Y_{t+1} = \kappa Y_t + \varepsilon_{t+1}, \text{ for } t \in \mathbb{Z},$$

where ε_t is standard normal for $t \in \mathbb{Z}$ and $|\kappa| < 1$. Clearly, $Y_t = \sum_{k=0}^{\infty} \kappa^k \varepsilon_{t-k}$, and therefore $Y_t \sim N\left(0, \frac{1}{1-\kappa^2}\right)$. For the sequences X_t and X'_t we have two options: either



(a) X_{k+1} and X'_{k+1} independent

(b) $X_{k+1} = X'_{k+1}$

Figure 3.4: Log-log plot of $\mathbb{E}|\theta^* - \theta_k|$ vs. number of iterations for i.i.d. beta innovations

we take consecutive measurements i.e. $X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$ or we use identical values, i.e. $X_k = X'_k = Y_k$. In both cases

$$U_4(\theta) = \mathbb{E}J(\theta, X) = \theta^2 + \frac{1}{1-\kappa^2} + \Phi\left(\theta\sqrt{1-\kappa^2}\right).$$

Solving this for $\kappa = 0.75$ we get the optimal value $\theta^* \approx -0.13144$.

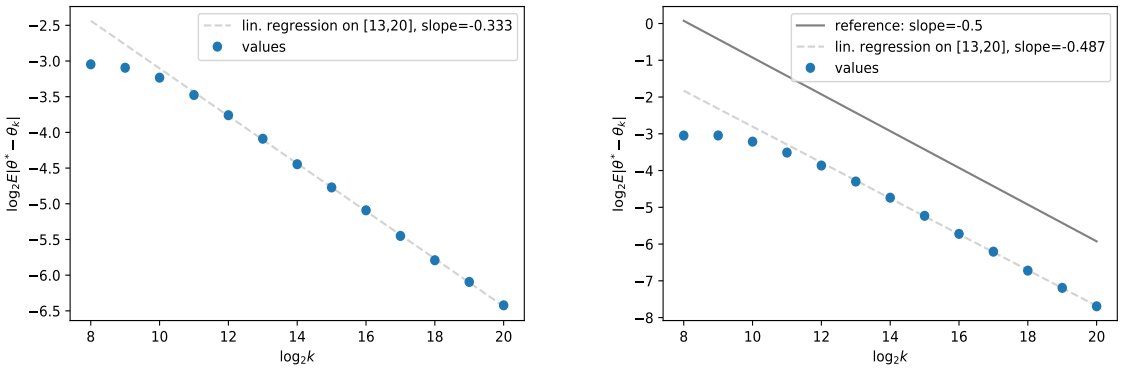
Figure 3.5 and table 3.2 illustrate the convergence rate of algorithm (3.19) for the function U_4, starting from $\theta_0 = 0$. On figure (3.5a) we present the rate in the case where we take consecutive measurements of the AR(1) process, ($X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$), the convergence rate of $k^{-0.333}$ was observed. Figure (3.5b) shows the case where the two measurements are the same, ($X_k = X'_k = Y_k$), with the rate $k^{-0.487}$.

	consecutive observations: $X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$	identical: $X_k = X'_k = Y_k$
AR(1)	-0.333 ($R^2 = 0.999$)	-0.487 ($R^2 = 0.999$)

Table 3.2: Convergence rate for AR(1) noise

3.6 Application to mathematical finance

The price of a financial asset either follows a trend during a given period of time or just rambles around its "fair" price value – at least so it seems to many actual traders. This "rambling", in more mathematical terms, means that the price is reverting to its long-term average. Such a mean-reversion phenomenon can be exploited by "buying low, selling high"-type strategies. Related discussions involve plenty of common-sense advice and benevolent concrete suggestion, see e.g. [TraderGav, 2020, WarriorTrading,



(a) $X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$ (b) $X_k = X'_k = Y_k$

Figure 3.5: Log-log plot of $\mathbb{E}|\theta^* - \theta_k|$ vs. number of iterations for AR(1) innovations

2020, TradingStrategyGuides, 2021]. There exist also theoretical studies about optimal trading with such prices, see e.g. [Guasoni et al., 2019]. However, a rigorous approach to *adaptive* trading algorithms of this type is lacking.

Results of the present paper provide theoretical convergence guarantees for such algorithms which cannot be deduced from existing literature on stochastic approximation. The most conspicuous feature of mean-reversion strategies is that they are triggered when the price reaches a certain level. This means that their payoffs are *discontinuous* with respect to the parameters, gradients do not exist and only finite-difference approximations can be used (the Kiefer-Wolfowitz method). Their convergence in the given discontinuous case cannot be shown based on available results hence we fill an important and practically relevant gap here.

We describe in some detail a trading model below and explain how it fits into the framework used in the previous sections. Let the price of the observed financial asset be described by a real-valued stochastic process S_t, $t \in \mathbb{Z}$, adapted to a given filtration $\mathcal{F}_t$, $t \in \mathbb{Z}$, representing the flow of information. (Alternatively S_t may be the *increment* of the price at t which can safely be assumed to follow a stationary process.)

Our algorithm will be based on several dynamically updated estimators which are assumed to be functionals of the trajectories of S_t and possibly of another adapted process F_t describing important economic factors. The estimate for the long-term average of the process is denoted by $A_t(\theta)$ at time t. The upper and lower bandwith processes will be denoted $B_t^+(\theta)$ and $B_t^-(\theta)$, they are non-negative. All these estimates depend on a parameter θ to be tuned, where θ ranges over a subset Q of $\mathbb{R}^d$.

In practice, $A_t(\theta)$ is some moving average (or exponential moving average) of previous values of S but it may depend on the other indicators F (market indices, etc.).

Here θ determines, for instance, the weights of the moving average estimate. The quantities $B_t^{\pm}(\theta)$ are normally based on standard deviation estimates for S but, again, may be more complex with θ describing weighting of past information. If we peek from time t back to time $t-p$ with some $p \in \mathbb{N}$ then $A_t(\theta), B_t^{\pm}(\theta)$ are functionals of $(S_{t-p}, F_{t-p}, \ldots, S_t, F_t)$.

The price range $[A_t - B_t^-, A_t + B_t^+]$ is considered to be "normal" by the algorithm while quitting that interval suggest "extremal" behaviour that the market should correct soon. For example, reaching the level $A_t - B_t^-$ means that the price is abnormally low for the present circumstances, hence it is worth buying a quantity $b(\theta)$ of them where, again, the parameter θ should be optimally found. When the price returns to $A_{t'}$ at some later time t', the asset will be sold and a profit is (hopefully) realized. Similarly, when reaching $A_t + B_t^+$, quantity $s(\theta)$ of the asset is sold (the price being abnormally high) and it will be repurchased once the "normal" level $A_{t'}$ is reached at some future $t' > t$, aiming to realize profit.

The value of the parameter θ will be updated at times tN, $t \in \mathbb{N}$ where $N \geq 1$ is fixed. The (random) profit (or loss) resulting from trading on the interval $[N(t-1), Nt]$ is denoted by $u(\theta, X_t)$ with $X_t = (S_{N(t-1)-p}, F_{N(t-1)-p}, \ldots, S_{Nt}, F_{Nt})$. We could even write an explicit expression for u based on the description of the trading mechanism in the previous paragraph but it would be very cumbersome without providing additional insight hence we omit it. We also add that, in many cases, a fee must also be paid at every transaction. Such strategies being "threshold-type", the function u is generically a *discontinuous* function of θ.

We furthermore argue that one *cannot* smooth out u and make it continuous without losing *essential* features of the problem. Approximating the indicator function of the interval $[0, \infty)$ by a function f which is 1 on $[0, \infty)$, 0 on $(-\infty, -\epsilon]$ for some small $\epsilon > 0$ and linear on $(-\epsilon, 0)$ may look reasonable at first sight but in this way we get a Lipschitz approximation with a huge Lipschitz constant hence with a poor convergence rate! This is just to stress that such simple tricks might work in certain practical situations but they only obscure the real issues in the theoretical analysis (namely, there *is* a discontinuity to be handled).

The described algorithm is very close to what actual investors do, see [TraderGav, 2020, WarriorTrading, 2020, TradingStrategyGuides, 2021]. We also mention the related theoretical studies [Leung and Li, 2015, Cartea et al., 2015] which, however, do not take an adaptive view and calculate optimal strategies for concrete models.

Taking a more realistic, adaptive approach, the investor may seek to maximize $Eu(\theta, X_0)$ by dynamically updating θ at every instant tN, $t \in \mathbb{N}$. Our versions of the Kiefer-Wolfowitz algorithm, presented in the previous sections, are tailor-made for

such online optimization, both the decreasing and the fixed gain version, depending on the circumstances. Theorems 3.2.1 and 3.2.2 provide a solid theoretical convergence guarantee for such procedures.

Chapter 4

On the stability of the stochastic gradient Langevin algorithm with dependent data stream

This chapter is based on the preprint [Rásonyi and Tikosi, 2021]. We prove, under mild conditions, that the fixed gain stochastic gradient Langevin dynamics converge to a limiting law as time tends to infinity, even in the case where the driving data sequence is dependent.

4.1 Stochastic gradient Langevin dynamics

Sampling from high-dimensional, possibly not even logconcave distributions is a challenging task, with far-reaching applications in optimization, in particular, in machine learning, see [Raginsky et al., 2017, Chau et al., 2021, Barkhagen et al., 2021, Brosse et al., 2018].

Let $U : \mathbb{R}^d \to \mathbb{R}_+$ be a given function and consider the corresponding Langevin equation

$$d\Theta_t = -\nabla U(\Theta_t)\, dt + \sqrt{2}\, dW_t, \tag{4.1}$$

where W is a d-dimensional standard Brownian motion. Under suitable assumptions, the unique invariant probability μ for the diffusion process (4.1) has a density (with respect to the d-dimensional Lebesgue measure) that is proportional to $\exp(-U(x))$, $x \in \mathbb{R}^d$.

In practice, Euler approximations of (4.1) may be used for sampling from μ, i.e. a recursive scheme

$$\vartheta^\lambda_{t+1} = \vartheta^\lambda_t - \lambda \nabla U(\vartheta^\lambda_t) + \sqrt{2\lambda}\xi_{t+1} \tag{4.2}$$

is considered for some small $\lambda > 0$ and independent standard d-dimensional Gaussian sequence ξ_i, $i \geq 1$.

In some important applications however, U and ∇U are unknown, one disposes only of unbiased estimates $H(\theta, Y_t)$, $t \in \mathbb{N}$ of $\nabla U(\theta)$, where Y_t is some stationary data sequence. From this point on we switch to rigorous mathematics.

Let us fix integers $d, m \geq 1$ and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $\mathcal{B}(\mathcal{X})$ denotes the σ-algebra of the Borel-sets of a Polish space[1] $\mathcal{X}$. For a random variable X, $\mathcal{L}(X)$ denotes its law. The Euclidean norm on $\mathbb{R}^d$ or $\mathbb{R}^m$ will be denoted by $|\cdot|$, while $||\cdot||_{TV}$ stands for the total variation distance of probability measures [2] on $\mathcal{B}(\mathbb{R}^d)$. Let $B_r := \{\theta \in \mathbb{R}^k : |\theta| \leq r\}$ denote the ball of radius r, for $r \geq 0$, for both $k = d$ and $k = m$, depending on the context. The notation $\mathrm{Leb}(\cdot)$ refers to the d-dimensional Lebesque-measure.

For $0 < \lambda \leq 1$, $t = 0, 1, \ldots$ and for a constant initial value $\theta_0 \in \mathbb{R}^d$ consider the recursion

$$\theta^\lambda_{t+1} = \theta^\lambda_t - \lambda H(\theta^\lambda_t, Y_t) + \sqrt{2\lambda}\xi_{t+1}, \ t \in \mathbb{N}, \ \theta^\lambda_0 := \theta_0, \tag{4.3}$$

where ξ_i, $i \geq 1$ is an i.i.d. sequence of d-dimensional random variables with independent coordinates such that $\mathbb{E}[\xi_i] = 0$ and $E[|\xi_i|^2] = \sigma^2$ for some $\sigma \geq 0$. Furthermore, the density function f of ξ_i with respect to the Lebesque-measure is assumed strictly positive on every compact set. Assume that $(Y_t)_{t\in\mathbb{Z}}$ is a strict sense stationary[3] process with values in $\mathbb{R}^m$ and it is independent of the noise process $(\xi_t)_{t\geq 1}$. Finally, $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a measurable function.

A particular case of (4.3) is the *stochastic gradient Langevin dynamics* (SGLD), introduced in [Welling and Teh, 2011], designed to learn from large datasets. See more about different versions of SGLD and their connections in [Brosse et al., 2018]. Note that in the present setting, unlike in SGLD, we do not assume that H is the gradient of a function and we do not assume ξ_i to be Gaussian.

A setting similar to ours was considered in [Lovas and Rásonyi, 2021] under different assumptions. We will compare our results to those of [Lovas and Rásonyi, 2021] at the end of Section 4.2 below.

The sampling error of θ^λ_t has been thoroughly analysed in the literature: $d(\mathcal{L}(\theta^\lambda_t), \mu)$ has been estimated for various probability metrics d, see [Chau et al., 2021, Barkhagen

[1] separable completely metrizable topological space

[2] The total variation distance of probability measures P and Q on a probability space $(\Omega, \mathcal{F})$ is defined as $||P - Q||_{TV} = \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$.

[3] The process $(x_k)_{k\in\mathbb{Z}}$ is called strongly stationary (or strictly stationary) if the distribution is time invariant, i.e. the joint distribution of $(x_{t_1}, \ldots x_{t_k})$ is the same as of $(x_{t_1+j}, \ldots x_{t_k+j})$ for every $t_1, \ldots, t_k$ indices and for all k and j.

et al., 2021, Raginsky et al., 2017, Brosse et al., 2018]. The ergodic behaviour of θ_t^λ, however, has eluded attention so far. If Y_t are i.i.d. then θ_t^λ is a homogeneous Markov chain and standard results of Markov chain theory apply. In the more general stationary case (considered in [Barkhagen et al., 2021, Chau et al., 2021]) however, that machinery is not available. In the present note we study scheme (4.3) with stationary Y_t and establish that its law converges to a limit in total variation.

4.2 Assumptions and main result

Below we state the assumptions for our main result. The first two assumptions concern the growth of the estimates $H(\theta, y)$. The following assumption is often referred to as *dissipativity*.

Assumption 4.2.1 *There is a constant $\Delta > 0$ and a measurable function $b : \mathbb{R}^m \to \mathbb{R}_+$ such that, for all $\theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$*

$$\langle H(\theta, y), \theta \rangle \geq \Delta |\theta|^2 - b(y). \tag{4.4}$$

Assumption 4.2.2 *There exist constants $K_1, K_2, K_3 > 0$ and $\beta \geq 1$ such that*

$$|H(\theta, y)| \leq K_1 |\theta| + K_2 |y|^\beta + K_3. \tag{4.5}$$

Assumption 4.2.2 is satisfied in particular if the function $H(\theta, y)$ is Lipschitz continuous in θ and has polynomial growth in y. The following assumption is posed on the elements of the stationary noise sequence Y.

Assumption 4.2.3 *There exist constants $M_y, M_b > 0$ such that $\mathbb{E}[|Y_0|^{2\beta}] \leq M_y$ and $\mathbb{E}[b(Y_0)] \leq M_b$.*

Theorem 4.2.1 *Let Assumptions 4.2.1, 4.2.2 and 4.2.3 hold. Then, for λ small enough, the law $\mathcal{L}(\theta_t^\lambda)$ of the iteration defined by (4.3) converges in total variation as $t \to \infty$ and the limit does not depend on the initialization X_0.*

In [Lovas and Rásonyi, 2021], Δ in (4.4) was allowed to depend on y but b in (4.4) had to be constant, the process Y was assumed bounded and the process ξ Gaussian. Furthermore, in Assumption 4.2.2, β had to be 1. Under these conditions the conclusion of Theorem 4.2.1 was obtained, together with a rate estimate.

Theorem 4.2.1 above complements the results of [Lovas and Rásonyi, 2021]: Δ must be constant in our setting but the restrictive boundedness hypothesis on Y could be removed, ξ need not be Gaussian, β in (4.5) can be arbitrary and b in (4.4) may depend on y. The examples in Section 4.5 demonstrate that our present results cover a wide range of relevant applications where the obtained generalizations are crucial.

4.3 Markov chains in random environment

The rather abstract Theorem 4.3.1 below, taken from [Gerencsér and Rásonyi, 2020], is the key result we use in this paper. Let us first recall the related terminology and the assumptions.

Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $(\mathcal{X}_n)_{n\in\mathbb{N}}$ (resp. $(\mathcal{Y}_n)_{n\in\mathbb{N}}$) be a non-decreasing sequence of (non-empty) Borel-sets in $\mathcal{X}$ (resp. $\mathcal{Y}$). Consider a parametric family of transition kernels, i.e. a map $Q : \mathcal{X} \times \mathcal{Y} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ such that for all $B \in \mathcal{B}(\mathcal{X})$ the function $(x,y) \to Q(x,y,B)$ is measurable and for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ $Q(x,y,\cdot)$ is a probability measure. The parameters y will play the role of the environment that affects the process $(X_t)_{t\in\mathbb{N}}$ defined below. Let Y_t be a $\mathcal{Y}$-valued strongly stationary process.

Definition 4.3.1 An $\mathcal{X}$ valued stochastic process $(X_t)_{t\in\mathbb{N}}$ is called a *Markov chain in a random environment* with transition kernel Q if $X_0 \in \mathcal{X}$ is deterministic (for simplicity) and

$$\mathbb{P}(X_{t+1} \in A|\mathcal{F}_t) = Q(X_t, Y_t, A), \text{ for } t \in \mathbb{N}, \tag{4.6}$$

where we use the filtration $\mathcal{F}_t = \sigma(Y_k, k \in \mathbb{Z}; X_j, 0 \le j \le t)$.

Definition 4.3.2 (kernels as operators) For a parametric family of transition kernels Q and a bounded (or non-negative) function $V : \mathcal{X} \to \mathbb{R}$ define

$$[Q(y)V](x) = \int_{\mathcal{X}} V(z)Q(x,y,dz), \text{ for } x \in \mathcal{X}. \tag{4.7}$$

This is in fact associating the kernel with the linear operator by the above definition.

We will use the short notation $Q(y)$ for the kernel $Q(\cdot, y, \cdot)$ and thus $Q(y)V$ means the action of $Q(y)$ on the function V. We recall the definition of the product of kernels.

Definition 4.3.3 The product of two kernels $Q(y_1)$ and $Q(y_2)$ is defined as

$$Q(y_2)Q(y_1)(x,B) = \int_{\mathcal{X}} Q(x,y_2,dz)Q(z,y_1,B).$$

The following tells that the starting point of the Markov chain is such, that the process fulfills a tightness-like assumption.

Assumption 4.3.1 *Let the process $(X_t)_{t\in\mathbb{N}}$ started from X_0 with be such that*

$$\sup_{t\in\mathbb{N}} \mathbb{P}(X_t \notin \mathcal{X}_n) \to 0, n \to \infty. \tag{4.8}$$

Assumption 4.3.2 *(Minorization condition) Let* $\mathbb{P}(Y_0 \notin \mathcal{Y}_n) \to 0, n \to \infty$. *Assume that there exists a sequence of probability measures* $(\nu_n)_{n\in\mathbb{N}}$ *and a non-decreasing sequence* $(\alpha_n)_{n\in\mathbb{N}}$ *with* $\alpha_n \in (0,1]$ *such that for all* $n \in \mathbb{N}, x \in \mathcal{X}_n, y \in \mathcal{Y}_n$, *and* $A \in \mathcal{B}(\mathcal{X})$,

$$Q(x, y, A) \geq \alpha_n \nu_n(A). \tag{4.9}$$

Here the lower bound $\alpha_n\nu_n(A)$ depends on x and y only though the sets $\mathcal{X}_n$ and $\mathcal{Y}_n$, thus this assumption tells, intuitively, that there exist sets A which have positive probability under the kernel regardless of the starting point x and the parameter y.

Theorem 4.3.1 *(Theorem 2.11. of [Gerencsér and Rásonyi, 2020]) Let Assumptions 4.3.1 and 4.3.2 hold. Then there exists a probability* μ_* *on* $\mathcal{B}(\mathcal{X} \times \mathcal{Y}^{\mathbb{Z}})$ *such that*

$$||\mathcal{L}(X_t, (Y_{t+k})_{k\in\mathbb{Z}}) - \mu_*||_{TV} \to 0, \text{ as } t \to \infty.$$

If $(X'_t)_{k\in\mathbb{N}}$ *is another such Markov chain started from a different* X'_0 *satisfying Assumption 4.3.1 then*

$$||\mathcal{L}(X_t, (Y_{t+k})_{k\in\mathbb{Z}}) - \mathcal{L}(X'_t, (Y_{t+k})_{k\in\mathbb{Z}})||_{TV} \to 0, \text{ as } t \to \infty. \quad \square$$

4.4 Proofs

At the end of this section we prove Theorem 4.2.1. To get there we will need certain lemmas.

Define the Markov chain associated to the recursive scheme (4.3) as

$$Q(\theta, y, A) = \mathbb{P}(\theta - \lambda H(\theta, y) + \sqrt{2\lambda}\xi_{n+1} \in A), \tag{4.10}$$

for all $y \in \mathcal{Y} := \mathbb{R}^m$, $\theta \in \mathcal{X} := \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$.

Lemma 4.4.1 *For small enough* λ, *under Assumptions 4.2.1 and 4.2.2, the process* $(\theta^\lambda_t)_{t\in\mathbb{N}}$ *given by recursion (4.3) satisfies Assumption 4.3.1 with* $\mathcal{X}_n := B_n$ *(the ball of radius* n*).*

Proof: Choose $V(\theta) = |\theta|^2$. Then, using that $E\xi_1 = 0$, $E[|\xi_i|^2] = \sigma^2$, Assumption 4.2.2 and Assumption 4.2.1, we get that

$$\begin{aligned}
[Q(y)V](\theta) &= \mathbb{E}[V(\theta - \lambda H(\theta, y) + \sqrt{2\lambda}\xi_1)] \\
&= |\theta|^2 + \lambda^2|H(\theta, y)|^2 + 2\lambda\mathbb{E}|\xi_1|^2 - 2\lambda\langle\theta, H(\theta, y)\rangle \\
&\leq (1 - 2\lambda\Delta)|\theta|^2 + 2\lambda(\sigma^2 + b(y)) + 3\lambda^2(K_1^2|\theta|^2 + K_2^2|y|^{2\beta} + K_3^2) \\
&= (1 - 2\lambda\Delta + 3\lambda^2K_1^2)V(\theta) + 2\lambda(\sigma^2 + b(y)) + 3\lambda^2(K_2^2|y|^{2\beta} + K_3^2) \\
&= \gamma V(\theta) + K(y),
\end{aligned}$$

with

$$K(y) = 2\lambda(\sigma^2 + b(y)) + 3\lambda^2[K_2^2|y|^{2\beta} + K_3^2]$$
$$\gamma = 1 - 2\lambda\Delta + 3\lambda^2 K_1^2.$$

Note that for small enough λ, $\gamma \in (0,1)$, independent of y.

Now using Lemma 4.4.2 below and setting $\theta = \theta_0$ and $y_k = Y_k$ for $k \geq 1$ we get, for each $t \geq 1$, that

$$\begin{aligned}
\mathbb{E}|\theta_t^\lambda|^2 &= \mathbb{E}[Q(Y_t)Q(Y_{t-1})\dots Q(Y_1)V](\theta_0) \\
&\leq \gamma^t V(\theta_0) + \sum_{i=1}^{t} \gamma^i \mathbb{E}K(Y_i) \\
&= \gamma^t|\theta_0|^2 + \sum_{i=1}^{t} \gamma^i[\lambda(\sigma^2 + 2\mathbb{E}[b(Y_i)]) + 3\lambda^2(K_2^2\mathbb{E}|Y_i|^{2\beta} + K_3^2)] \\
&\leq |\theta_0|^2 + \frac{\gamma}{1-\gamma}[(\sigma^2 + 2M_b) + 3(K_2^2 M_y + K_3^2)] < \infty,
\end{aligned}$$

by Assumption 4.2.3. Then, using Markov's inequality, we arrive at

$$\mathbb{P}(\theta_t^\lambda \notin \mathcal{X}_n) = \mathbb{P}(|\theta_t^\lambda| > n) \leq \frac{\sup_t \mathbb{E}|\theta_t^\lambda|^2}{n^2} \to 0, \text{ as } n \to \infty.$$

□

Lemma 4.4.2 *Assume* $[Q(y)V](\theta) \leq \gamma V(\theta) + K(y)$. *Then*

$$[Q(y_k)Q(y_{k-1})\dots Q(y_1)V](\theta) \leq \gamma^k V(\theta) + \sum_{i=1}^{k} \gamma^{i-1} K(y_i). \tag{4.11}$$

Proof: We prove the statement by induction. For $k = 1$, it is true by assumption. Combining definition 4.3.2 and 4.3.3 we get that

$$[Q(y_2)Q(y_1)V](x) = \int_{\mathcal{X}} Q(x, y_2, dr) \int_{\mathcal{X}} V(z)Q(r, y_1, dz), \text{ for } r \in \mathcal{X},$$

and then for $k > 1$

$$\begin{aligned}
[Q(y_k)Q(y_{k-1})\dots Q(y_1)V](\theta) &= \int_{\mathcal{X}} Q(\theta, y_k, dx)[Q(y_{k-1})Q(y_{k-2})\dots Q(y_1)V](x) \\
&\leq \int_{\mathcal{X}} \left(\gamma^{k-1}V(x) + \sum_{i=1}^{k-1} \gamma^{i-1}K(y_i)\right) Q(\theta, y_k, dx) \\
&= \gamma^{k-1} \int_{\mathcal{X}} V(x)Q(\theta, y_k, dx) + \sum_{i=1}^{k-1} \gamma^{i-1}K(y_i) \\
&\leq \gamma^k V(\theta) + \sum_{i=1}^{k} \gamma^{i-1}K(y_i).
\end{aligned}$$

□

Lemma 4.4.3 *Define $\mathcal{X}_n = B_n$, $\mathcal{Y}_n := B_n$, $n \in \mathbb{N}$ and let Assumptions 4.2.1 and 4.2.2 hold. Then Assumption 4.3.2 is satisfied, for all λ.*

Proof: For all $A \in \mathcal{B}(\mathcal{X})$,

$$\begin{aligned} Q(\theta, y, A) &= \mathbb{P}(\theta - \lambda H(\theta, y) + \sqrt{2\lambda}\xi_1 \in A) \\ &\geq \int_{\mathbb{R}^d} \mathbb{1}_{\{\theta - \lambda H(\theta,y) + \sqrt{2\lambda}\xi_1 \in A \cap B_n\}} f(w) dw \\ &= \frac{1}{\lambda^{d/2}} \int_{A \cap B_n} f\left(\frac{z - \theta + \lambda H(\theta, y)}{\sqrt{2\lambda}}\right) dz \\ &\geq \frac{\text{Leb}(A \cap B_n)}{\lambda^{d/2}} C(n) \\ &= \frac{\text{Leb}(A \cap B_n)}{\text{Leb}(B_n)} \frac{C(n)\text{Leb}(B_n)}{\lambda^{d/2}}, \end{aligned}$$

where we use that for $\theta, z \in B_n$ and $y \in B_n$ we have

$$\left|\frac{z - \theta + \lambda H(\theta, y)}{\sqrt{2\lambda}}\right| \leq \frac{n + n + \lambda(K_1 n + K_2 n^\beta + K_3)}{\sqrt{2\lambda}} =: R(n),$$

therefore the integrand can be bounded from below by $C(n) := \inf_{x \in B_{R(n)}} f(x) > 0$. Then define

$$\nu_n(A) := \frac{\text{Leb}(A \cap B_n)}{\text{Leb}(B_n)} \text{ and } \alpha_n := \frac{C(n)\text{Leb}(B_n)}{\lambda^{d/2}},$$

which proves that Assumption 4.3.2 holds. □

Proof: [of Theorem 4.2.1.] Lemmas 4.4.1 and 4.4.3 ensure that the assumptions of Theorem 4.3.1 hold, from which the statement follows. □

4.5 Examples

4.5.1 Multiple minima

Below we present a simple example of an objective function with two minima and we check that it satisfies our assumptions.

Consider the function $J : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ defined below.

$$J(\theta, Y) = \begin{cases} (\theta - Y)^2 & \text{if } \theta \leq 1 \\ -(\theta - 2 - Y)^2 + 4 & \text{if } 1 \leq \theta \leq 3 \\ \frac{1}{2}(\theta - 5 - Y)^2 - 1 & \text{otherwise,} \end{cases}$$

where $Y \sim N(0, 1)$ (or any other distribution with 0 mean and finite second moment).

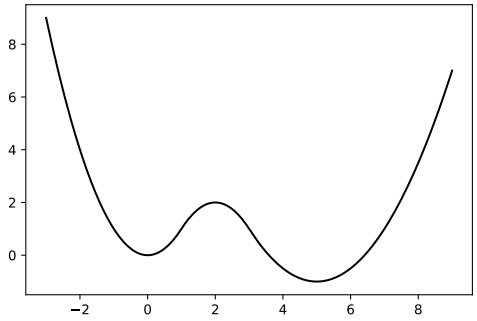


Figure 4.1: The function $U(\theta) = EJ(\theta, Y)$ with two minima

Then $U(\theta) = EJ(\theta, Y)$ (figure 4.1) has two minima in $\theta = 0$ and $\theta = 5$. $U(\theta)$ is differentiable and with $H(\theta, Y) = \frac{\partial J}{\partial \theta}(\theta, Y)$ we get that

$$H(\theta, Y) = \frac{\partial J}{\partial \theta}(\theta, Y) = \begin{cases} 2\theta - 2Y & \text{if } \theta \leq 1 \\ -2\theta + 2Y + 4 & \text{if } 1 \leq \theta \leq 3 \\ \theta - Y - 5 & \text{otherwise,} \end{cases}$$

$$< H(\theta, Y), \theta >= \begin{cases} 2|\theta|^2 - 2Y\theta & \text{if } \theta \leq 1 \\ -2|\theta|^2 + 2Y\theta + 4\theta & \text{if } 1 \leq \theta \leq 3 \\ |\theta|^2 - Y\theta - 5\theta & \text{otherwise.} \end{cases}$$

Then the function H satisfies Assumption 4.2.2 with $\beta = 1$ and Assumption 4.2.1 with $\Delta = 0.5$ and $b(Y) = Y^2 + 25$.

4.5.2 Nonlinear regression

Let us consider a nonlinear regression problem which can also be seen as a one layer neural network in a supervised learning setting, where only one trainable layer connects the input and the output vectors. The training set consists of entries $Y_t = (Z_t, L_t)$ with the features $Z_t \in \mathbb{R}^{d_0}$ and the corresponding labels $L_t \in \mathbb{R}^{d_1}$ for $t \in 1, \ldots, N$. We assume that Y_t is a stationary process. Set $m := d_0 + d_1$, the dimension of Y_t.

The trainable parameters will be a matrix $W \in \mathbb{R}^{d_0 \times d_1}$ and a vector $g \in \mathbb{R}^{d_1}$, therefore the dimension of $\theta := (W, g)$ will be $d = d_0 d_1 + d_1$. The prediction function $h : \mathbb{R}^{d_0} \times \mathbb{R}^d \to \mathbb{R}^{d_1}$ is defined by

$$h(z, \theta) := s(Wz + g),$$

where $s = (s_1, \ldots, s_{d_1})$ is a collection of nonlinear activation functions $s_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \ldots, d_1$. We will assume that each s_i and their derivatives s_i' are all bounded by some constant M_s for $i = 1, \ldots, d_1$.

Remark 4.5.1 Some widely used differentiable activation functions are bounded with bounded derivatives (sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, softsign $\frac{x}{1+|x|}$, etc.)

Choosing the loss function to be mean-square error, one aims to minimize the empirical risk, that is

$$\min\{\mathbb{E}[|h(Z_t, \theta) - L_t|^2] + \kappa|\theta|^2\}, \tag{4.12}$$

with some $\kappa > 0$, where the second term is added for regularization.

It is standard to solve this optimization step using gradient-based methods. For $y = (z, l) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$ denote

$$U(\theta, y) = |h(z, \theta) - l|^2 + \kappa|\theta|^2$$

and the updating function to be used in the algorithm will be

$$H(\theta, y) = \nabla U(\theta, y) = \frac{\partial}{\partial\theta}|h(z, \theta) - l|^2 + 2\kappa\theta. \tag{4.13}$$

Lemma 4.5.1 *The function $H(\theta, y)$ defined as above satisfies Assumptions 4.2.1 and 4.2.2.*

Proof: Using the chain rule, a short calculation gives

$$\left|\frac{\partial}{\partial\theta}|h(z;\theta) - l|^2\right| = \sqrt{\sum_{i=1}^{d_0+1}\sum_{j=1}^{d_1}\left(2(h(z;\theta)_j - l_j)s_j'(\langle W_j, z\rangle + g_j)z_i\right)^2}, \tag{4.14}$$

where we define $z_{d_0+1} = 1$ and W_j stands for the jth row of W. Notice that by the boundedness of s' and s this is at most quadratic in y. Then Assumption 4.2.2 is satisfied with $\beta = 2$.

Using the same argument about the boundedness of s and s'

$$\begin{aligned}
&\left|\left\langle\frac{\partial}{\partial\theta}|h(z;\theta) - l|^2, \theta\right\rangle\right| \\
=&\left|\sum_{i=1}^{d_0+1}\sum_{j=1}^{d_1} 2(h(z;\theta)_j - l_j)s_j'(W_j z + g_j)z_i\theta_{i,j}\right| \\
\leq& C_0\left(|y|^2 + 1\right)|\theta|,
\end{aligned}$$

for some $C_0 > 0$. Using that $\left\langle\frac{\partial}{\partial\theta}\kappa|\theta|^2, \theta\right\rangle = 2\kappa|\theta|^2$, we get that $\langle\nabla U(\theta), \theta\rangle \geq c|\theta|^2 - C(|y|^4 + 1)$ with some c, C therefore Assumption 4.2.1 is satisfied with $b(y)$ being of degree 4 in y. □

4.5.3 A tamed algorithm for neural networks

It has been observed that in multi-layer neural networks quadratic regularization is not always sufficient to guarantee convergence of the SGLD scheme, while adding a higher order term would violate Lipschitz continuity. So the standard SGLD algorithm diverges anyway. To remedy this, certain "tamed" schemes have been suggested in [Lovas et al., 2021].

In contrast to the previous case now we will have hidden layers between the input and output: layer 0 is the input, layer n is the output and $1, \dots, n-1$ are the hidden layers of the neural network for some $n > 1$. The prediction function h will be defined as the composition of a sequence of $n+1$ linear transformations and activation functions, i.e.

$$h(z, \theta) = s_n(W_n s_{n-1}(W_{n-1} \dots s_0(W_0 z))),$$

where θ is the collection of all parameters $W_i \in \mathbb{R}^{d_{i-1}} \times \mathbb{R}^{d_i}$, $i = 1, \dots, n$ and $s_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ is a componentwise non-linear activation function, assumed bounded together with its derivatives by some constant M_s. Therefore $h : \mathbb{R}^{d_0} \times \mathbb{R}^d \to \mathbb{R}^{d_n}$, where $d = \sum_{i=1}^n d_{i-1} d_i$ is the dimension of θ. For the case of simplicity in this case we assumed that there is no bias term g as in the previous example. The training set consists of entries $Y_t = (Z_t, L_t)$ with the features $Z_t \in \mathbb{R}^{d_0}$ and the corresponding labels $L_t \in \mathbb{R}^{d_n}$, the dimension of each Y_t is $m = d_0 + d_n$. We assume that Y_t is a stationary process.

As in the previous subsection, the regularized empirical risk has the form

$$U(\theta, y) = |h(z, \theta) - l|^2 + \frac{\eta}{2(r+1)} |\theta|^{2(r+1)}$$

with some $r \geq 0$, $\eta > 0$. Denoting $G(\theta, y) = \nabla U(\theta, y)$, the "tamed" updating function we use will be defined as

$$H(\theta, y) := \frac{G(\theta, y)}{1 + \sqrt{\lambda} |\theta|^{2r}}, \text{ for every } \theta \in \mathbb{R}^d, y \in \mathbb{R}^m.$$

Note that this function depends on λ!

We will use the following.

Lemma 4.5.2 *(Proposition 4 of [Lovas et al., 2021])*

$$\left| \frac{\partial}{\partial \theta} |h(z, \theta) - l|^2 \right| \leq C(1 + |y|)^2 \left(1 + |\theta|^{n+1}\right), \tag{4.15}$$

where $C > 0$ depends on $D = \max_{j=1,\dots,n} d_j$, n and M_s. □

Lemma 4.5.3 *For $\lambda < 1$ and η small enough, the conclusions of Theorem 4.2.1 hold for the scheme (4.3) with $H(\theta, y)$ defined as above, provided that $r \geq \frac{n+2}{2}$ and $\mathbb{E}[|Y_0|^2] < \infty$.*

Proof: Using Lemma 4.5.2, Assumption 4.2.2 can be checked as follows:

$$\begin{aligned}|H(\theta,y)| &= \left|\frac{\frac{\partial}{\partial\theta}|h(z;\theta)-l|^2+\eta\theta|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right| \\ &\leq \left|\frac{C(1+|y|)^2\,(1+|\theta|^{n+1})}{1+\sqrt{\lambda}|\theta|^{2r}}\right| + \left|\frac{\eta\theta|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right| \\ &\leq K_1|\theta| + K_2|y|^{\beta} + K_3,\end{aligned}$$

where $K_1 = \frac{\eta}{\sqrt{\lambda}}$, $\beta = 2$ and the constants K_2 and K_3 depend on λ, η, n and C.

Let us check Assumption 4.2.1. For the regularization term we have

$$\left\langle \frac{\eta\theta|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}, \theta \right\rangle = \frac{\eta|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}} \geq \min\left\{\frac{\eta}{2\sqrt{\lambda}}, \frac{\eta}{2}\right\}|\theta|^2 \geq \frac{\eta}{2}|\theta|^2 \tag{4.16}$$

for $\lambda < 1$.

The Cauchy inequality, Lemma 4.5.2 and the choice of r ensure that

$$\left|\left\langle \frac{\frac{\partial}{\partial\theta}\left(|h(z;\theta)-l|^2\right)}{1+\sqrt{\lambda}|\theta|^{2r}}, \theta \right\rangle\right| \leq \frac{C'(1+|\theta|^{n+2})(1+|y|^2)}{1+\sqrt{\lambda}|\theta|^{2r}} \leq K'(1+|y|^2), \tag{4.17}$$

for some $C', K' > 0$. Now combining these estimates, we get

$$\langle H(\theta,y),\theta\rangle \geq \frac{\eta}{2}|\theta|^2 - K'(1+|y|^2), \tag{4.18}$$

therefore Assumption 4.2.1 is satisfied with $\Delta = \frac{\eta}{2}$ and $b(y)$ is quadratic in y.

We can check that $\gamma = (1-\eta\lambda+3\lambda\eta^2) < 1$ in Lemma 4.4.1 for η small enough so the proof of Theorem 4.2.1 goes through for this choice of H. □

Assuming Y_0 to have a finite moment of order 4, $\frac{n+2}{2}$ in Lemma 4.5.3 could be decreased to $\frac{n+1}{2}$, as easily seen.

Appendix A

Notes on numerical experiments

A.1 Monte Carlo simulation

To numerically compute $\mathbb{E}|\theta_n - \theta^*|$ for a given starting point θ_0 one can use Monte Carlo simulation. In this case that means setting a number m of trials and running the algorithm m times for n steps. Then we approximate the expected value with the average of the simulations i.e.

$$\mathbb{E}|\theta_n - \theta^*| \approx \frac{1}{m}\sum_{i=1}^{m} \theta_n^{(m)},$$

where $\theta_n^{(m)}$ denotes the n^{th} recursion step of the m^{th} simulation.

A.2 Log-log plots

When numerically measuring convergence speed we used **log-log plots**. The idea is that we assume that the converge speed can be described as

$$\mathbb{E}|\theta_k - \theta^*| = ck^{\beta},$$

where c is some constant and β is the convergence rate that we are looking for. Such a function will be a straight line on a log-log plot (the x axis being $\log k$ and the y axis being $\log \mathbb{E}|\theta_k - \theta^*|$), as

$$\log \mathbb{E}|\theta_k - \theta^*| = \log c + \beta \log k.$$

Then the task translates to finding the slope of this straight line that fit to the data points.

A.3 Linear regression and the goodness of fit

In the numerical experiments in Chapter 3.5 the line we fits best to the data is the simple least squares linear regression line: for data points $\{x_1, y_1, \ldots x_n, y_n\}$ find α, β such that $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ is minimal.

One way to quantify the goodness of the fit is to use the R^2 value, which intuitively describes how much of the variance is explained and is defined as

$$R^2 = \frac{\text{sum of squared residuals}}{\text{sum of total residuals}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2},$$

where $\hat{y}_i$ are the fitted values and $\bar{y}$ is the mean of the data points.

Bibliography

[Ascher and Petzold, 1998] Ascher, U. M. and Petzold, L. R. (1998). *Computer methods for ordinary differential equations and differential-algebraic equations*, volume 61. Siam. 30

[Bardou et al., 2009] Bardou, O., Frikha, N., and Pages, G. (2009). Computing var and cvar using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods and Applications*, 15(3):173–210. 21

[Barkhagen et al., 2019] Barkhagen, M., Chau, N. H., Moulines, É., Rásonyi, M., Sabanis, S., and Zhang, Y. (2019). On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *To appear in Bernoulli, arXiv:1812.02709*. 45

[Barkhagen et al., 2021] Barkhagen, M., Chau, N. H., Moulines, Éric., Rásonyi, M., Sabanis, S., and Zhang, Y. (2021). On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27:1–33. 13, 54, 55, 56

[Benveniste et al., 1990] Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*. Springer. 27

[Benveniste and Ruget, 1982] Benveniste, A. and Ruget, G. (1982). A measure of the tracking capability of recursive stochastic algorithms with constant gains. *IEEE Transactions on Automatic Control*, 27(3):639–649. 9

[Bhatnagar et al., 2013] Bhatnagar, S., Prasad, H., and Prashanth, L. (2013). Stochastic approximation algorithms. In *Stochastic Recursive Algorithms for Optimization*, pages 17–28. Springer. 8, 9

[Blum et al., 1954] Blum, J. R. et al. (1954). Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386. 10

[Bottou et al., 2018] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311. 14

[Brosse et al., 2018] Brosse, N., Durmus, A., and Moulines, E. (2018). The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278. 11, 13, 54, 55

[Cartea et al., 2015] Cartea, Á., Jaimungal, S., and Penalva, J. (2015). *Algorithmic and high-frequency trading*. Cambridge University Press. 52

[Cauchy et al., 1847] Cauchy, A. et al. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538. 4

[Chau et al., 2019a] Chau, H. N., Kumar, C., Rásonyi, M., and Sabanis, S. (2019a). On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23:217–244. 9, 25, 28, 29, 31, 45, 46

[Chau et al., 2019b] Chau, N. H., Moulines, É., Rásonyi, M., Sabanis, S., and Zhang, Y. (2019b). On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *Preprint, arXiv:1905.13142*. 27

[Chau et al., 2021] Chau, N. H., Moulines, Éric., Rásonyi, M., Sabanis, S., and Zhang, Y. (2021). On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *Preprint, arXiv:1905.13142*. 13, 54, 55, 56

[Chung, 1954] Chung, K. L. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483. 8

[Duflo, 2013] Duflo, M. (2013). *Random iterative models*, volume 34. Springer Science & Business Media. 6, 7

[Durmus and Moulines, 2017] Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27:1551–1587. 34

[Fort et al., 2016] Fort, G., Moulines, É., Schreck, A., and Vihola, M. (2016). Convergence of markovian stochastic approximation with discontinuous dynamics. *SIAM Journal on Control and Optimization*, 54(2):866–893. 25

[Gerencsér and Rásonyi, 2020] Gerencsér, B. and Rásonyi, M. (2020). Invariant measures for fractional stochastic volatility models. *Preprint, arXiv:2002.04832v1*. 57, 58

[Gerencsér, 1989] Gerencsér, L. (1989). On a class of mixing processes. *Stochastics*, 26(3):165–191. 29

[Gerencsér, 1992] Gerencsér, L. (1992). Rate of convergence of recursive estimators. *SIAM Journal on Control and Optimization*, 30(5):1200–1227. 27, 30

[Gerencsér, 1998] Gerencsér, L. (1998). Spsa with state-dependent noise-a tool for direct adaptive control. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, volume 3, pages 3451–3456. IEEE. 30

[Gerencsér, 1999] Gerencsér, L. (1999). Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Transactions on Automatic Control*, 44(5):894–905. 30

[Gerencsér et al., 2007] Gerencsér, L., Vágó, Z., and Hill, S. D. (2007). The magic of spsa. In *PAMM: Proceedings in Applied Mathematics and Mechanics*, volume 7, pages 1062501–1062502. Wiley Online Library. 11

[Glasserman and Yao, 1992] Glasserman, P. and Yao, D. D. (1992). Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908. 47

[Guasoni et al., 2019] Guasoni, P., Tolomeo, A., and Wang, G. (2019). Should commodity investors follow commodities' prices? *SIAM Journal on Financial Mathematics*, 10(2):466–490. 51

[Kiefer and Wolfowitz, 1952] Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466. 6, 9, 25, 29

[Kushner and Clark, 2012] Kushner, H. J. and Clark, D. S. (2012). *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media. 30

[Kushner and Huang, 1981] Kushner, H. J. and Huang, H. (1981). Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105. 9

[Laruelle and Pagès, 2012] Laruelle, S. and Pagès, G. (2012). Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Methods and Applications*, 18(1):1–51. 21, 22, 25

[Leung and Li, 2015] Leung, T. and Li, X. (2015). Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical and Applied Finance*, 18(03):1550020. 52

[Ljung, 1977] Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22(4):551–575. 8, 29

[Ljung, 1980] Ljung, L. (1980). The ode approach to the analysis of adaptive control systems?? possibilities and limitations. In *Joint Automatic Control Conference*, number 17, page 10. 8

[Lovas et al., 2021] Lovas, A., Lytras, I., Rásonyi, M., and Sabanis, S. (2021). Taming neural networks with TUSLA: Non-convex learning via adaptive stochastic gradient Langevin algorithms. *Preprint, arXiv:2006.14514*. 63

[Lovas and Rásonyi, 2021] Lovas, A. and Rásonyi, M. (2021). Markov chains in random environment with applications in queueing theory and machine learning. *To appear in Stochastic Processes and their Applications, arXiv:1911.04377*. 55, 56

[Nemirovskij and Yudin, 1983] Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization. 8

[Polyak, 1987] Polyak, B. T. (1987). Introduction to optimization. *Optimization Software, Inc, New York*. 6, 7

[Raginsky et al., 2017] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Nonconvex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *Proceedings of Machine Learning Research*, 65:1674–1703. 11, 13, 54, 55

[Rásonyi and Tikosi, 2020] Rásonyi, M. and Tikosi, K. (2020). Convergence of the kiefer-wolfowitz algorithm in the presence of discontinuities. *arXiv preprint arXiv:2007.14069*. 4, 24

[Rásonyi and Tikosi, 2021] Rásonyi, M. and Tikosi, K. (2021). On the stability of the stochastic gradient langevin algorithm with dependent data stream. *arXiv preprint arXiv:2105.01422*. 5, 54

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407. 6, 7, 25

[Sacks, 1958] Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405. 30

[Spall, 2005] Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons. 11, 47

[Spall et al., 1992] Spall, J. C. et al. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341. 10

[Teh et al., 2016] Teh, Y. W., Thiery, A. H., and Vollmer, S. J. (2016). Consistency and fluctuations for stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17. 12

[TraderGav, 2020] TraderGav (2020). What is mean reversion trading strategy. https://tradergav.com/what-is-mean-reversion-trading-strategy/. 50, 52

[TradingStrategyGuides, 2021] TradingStrategyGuides (2021). Mean reversion trading strategy with a sneaky secret. https://tradingstrategyguides.com/mean-reversion-trading-strategy/. 50, 52

[Uryasev and Rockafellar, 2001] Uryasev, S. and Rockafellar, R. T. (2001). Conditional value-at-risk: optimization approach. In *Stochastic optimization: algorithms and applications*, pages 411–435. Springer. 21

[WarriorTrading, 2020] WarriorTrading (2020). Mean reversion trading: Is it a profitable strategy? https://www.warriortrading.com/mean-reversion/. 50, 52

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. 11, 12, 55

[Yin et al., 2002] Yin, G., Liu, R., and Zhang, Q. (2002). Recursive algorithms for stock liquidation: A stochastic optimization approach. *SIAM Journal on Optimization*, 13(1):240–263. 18, 20

[Zhang and Zhang, 2008] Zhang, H. and Zhang, Q. (2008). Trading a mean-reverting asset: Buy low and sell high. *Automatica*, 44(6):1511–1518. 16

[Zhuang, 2008] Zhuang, C. (2008). *Stochastic approximation methods and applications in financial optimization problems*. PhD thesis, University of Georgia. 16, 18, 20