Electron. J. Probab. **28** (2023), article no. 57, 1–48.
ISSN: 1083-6489 https://doi.org/10.1214/23-EJP947

Central limit theorems for stochastic gradient descent with averaging for stable manifolds*

Steffen Dereich† Sebastian Kassing‡

Abstract

In this article, we establish new central limit theorems for Ruppert-Polyak averaged stochastic gradient descent schemes. Compared to previous work we do not assume that convergence occurs to an isolated attractor but instead allow convergence to a stable manifold. On the stable manifold the target function is constant and the oscillations of the iterates in the tangential direction may be significantly larger than the ones in the normal direction. We still recover a central limit theorem for the averaged scheme in the normal direction with the same rates as in the case of isolated attractors. In the setting where the magnitude of the random perturbation is of constant order, our research covers step-sizes $\gamma_n = C_\gamma n^{-\gamma}$ with $C_\gamma > 0$ and $\gamma \in (\frac{3}{4}, 1)$. In particular, we show that the beneficial effect of averaging prevails in more general situations.

Keywords: stochastic approximation; Robbins-Monro; Ruppert-Polyak average; deep learning; stable manifold.
MSC2020 subject classifications: Primary 62L20, Secondary 60J05; 65C05.
Submitted to EJP on December 19, 2019, final version accepted on April 16, 2023.
Supersedes arXiv:1912.09187.

1 Introduction

We consider stochastic gradient descent (SGD) algorithms for the approximation of minima of functions $-F : \mathbb{R}^d \to \mathbb{R}$, where, at each point $x \in \mathbb{R}^d$, we are only able to simulate a noisy version of the gradient $f(x) = DF(x)$.

Stochastic approximation methods form a popular class of optimisation algorithms with applications in diverse areas of statistics, engineering and computer science. Nowadays, a key application lies in machine learning where it is used in the training of neural networks. The original concept was introduced 1951 by Robbins and Monro

*Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044–390685587, Mathematics Münster: Dynamics–Geometry–Structure.

†Institute for Mathematical Stochastics, Faculty of Mathematics and Computer Science, University of Münster, Germany. E-mail: steffen.dereich@uni-muenster.de

‡Faculty of Mathematics, University of Bielefeld, Germany. E-mail: skassing@math.uni-bielefeld.de

[33] and since then analysed in various directions, see e.g. [4, 13, 39, 14, 10, 34]. In previous research, a typical key assumptions is that $-F$ has *isolated* local minima with elliptic Hessian. Under this assumption, Sacks [36] proved a central limit theorem which showed that, for step-sizes $\gamma_n = C_\gamma n^{-1}$ with sufficiently large constant $C_\gamma > 0$, SGD converges to a minimum of the objective function with rate $n^{-1/2}$ which is typically the best convergence rate that can be obtained. However, this particular choice for the step-sizes has two disadvantages. First, the SGD typically needs too long (for practical applications) to get into the vicinity of a local minimum of $-F$ if the gradient flow itself needs a long time to approach a minimum. Second, whether a choice of the constant C_γ is appropriate depends on the ellipticity of the Hessian at the local minimum approached by the SGD and in practical applications the latter is typically not known. Therefore, one would like to work with larger step-sizes $\gamma_n = C_\gamma n^{-\gamma}$ with $C_\gamma > 0$ and $\gamma \in (\frac{1}{2}, 1)$. In that case, the convergence to a minimum is typically of order $n^{-\gamma/2}$, see e.g. [5, 21]. As found by Ruppert [35] and Polyak [30, 31], it is still possible to get convergence of order $n^{-1/2}$ by considering the running average of the iterates instead of the iterates themself. Following these original papers a variety of results were derived for the Robbins-Monro algorithm and the Ruppert-Polyak average and we refer the reader to the monographs [1, 29, 2, 25, 12, 22] for more details. We stress that previous research was focused on dynamical systems that converge to isolated minimisers of $-F$ and one of the main contributions of this article is to show that the beneficial effect of averaging prevails also in more general situations.

Classical convexity assumptions are often not met in practice and as an example we outline an application from machine learning [38, 24]. Many risk functions for the optimisation of the weights and biases of a neural network depend solely on the realisation function that the network generates. For neural networks with ReLU activation function the positive homogeneity of the activation function entails that every (representable) function possesses a non-discrete set of representations as deep learning network, see e.g. [9]. For smooth activation functions, Cooper [6] applied a variant of the implicit function theorem to show that, in an overparameterised setting, the set $M := \{x \in \mathbb{R}^d : f(x) = 0\}$ forms a lower-dimensional submanifold of $\mathbb{R}^d$.

It appears natural to ask for extensions on settings where the set of (local) minima forms a stable manifold. So far research in that direction is very limited. Fehrman et al. [16] established rates for the convergence of the target function of a stochastic gradient descent scheme under the assumption that the set of minima forms a stable manifold. We also mention the work by Tripuraneni et al. [37] where an averaging method for SGD on submanifolds is introduced so that the Ruppert-Polyak result is applicable for the approximation of an isolated stable minimum of a function defined on a Riemannian manifold.

Let us introduce the central dynamical system considered in this article. Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}_0}, \mathbb{P})$ be a filtered probability space and $F : \mathbb{R}^d \to \mathbb{R}$ a measurable and differentiable function and set $f = DF : \mathbb{R}^d \to \mathbb{R}^d$. Let M be a d_ζ-dimensional C^1-submanifold of $\mathbb{R}^d$ with

$$f\big|_M \equiv 0.$$

We consider an adapted dynamical system $(X_n)_{n\in\mathbb{N}_0}$ satisfying for all $n \in \mathbb{N}$

$$X_n = X_{n-1} + \gamma_n(f(X_{n-1}) + D_n), \tag{1.1}$$

where

(0) X_0 is a $\mathcal{F}_0$-measurable $\mathbb{R}^d$-valued random variable, the *starting value*,
(I) $(D_n)_{n\in\mathbb{N}}$ is an $\mathbb{R}^d$-valued, adapted process, the *perturbation*,
(II) $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals, the *step-sizes*.

We briefly refer to $(X_n)_{n\in\mathbb{N}_0}$ as the *Robbins-Monro system*. Furthermore, we consider for $n \in \mathbb{N}$ the *Ruppert-Polyak average with burn-in* given by

$$\bar{X}_n = \frac{1}{\bar{b}_n} \sum_{i=n_0(n)+1}^{n} b_i X_i, \tag{1.2}$$

where

(III) $(n_0(n))_{n\in\mathbb{N}}$ is a $\mathbb{N}_0$-valued sequence with $n_0(n) < n$ for all $n \in \mathbb{N}$ and $n_0(n) \to \infty$,
(IV) $(b_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals and $\bar{b}_n = \sum_{i=n_0(n)+1}^{n} b_i$ for $n \in \mathbb{N}$.

Roughly speaking, we raise and (at least partially) answer the following questions.

- Is Ruppert-Polyak averaging still beneficial in the case of non-isolated minimizers?
- If so, what are good choices for the parameters introduced in (II) to (IV)?

We answer these questions by deriving central limit theorems for the performance of the Ruppert-Polyak average on the event of convergence of $(X_n)_{n\in\mathbb{N}_0}$ to some element of the stable manifold M.

Let us be more precise. By assumption, M is a C^1-manifold and we will impose additional regularity assumptions on the tangent spaces (see Definition 2.4) that will guarantee existence of an open neighbourhood $\mathbb{M}$ of M so that for each $x \in \mathbb{M}$ there exists a unique closest element x^* in M, the *M-projection of x* (cf. [11], [23]). We denote by $\mathbb{M}^{\mathrm{conv}}$ the event that $(X_n)_{n\in\mathbb{N}_0}$ converges to an element of M and denote the limit by X_∞. Note that, on $\mathbb{M}^{\mathrm{conv}}$, the M-projection X_n^* and $\bar{X}_n^*$ are well-defined for sufficiently large (random) n and we will provide stable limit theorems for

$$\sqrt{n}(\bar{X}_n - \bar{X}_n^*) \text{ and } n\,(F(X_\infty) - F(\bar{X}_n)),$$

on the event $\mathbb{M}^{\mathrm{conv}}$. It is natural to consider Ruppert-Polyak averaging only in the case, where the Robbins-Monro scheme converges. In particular, this guarantees that the average converges to the same limit. Our analysis is conducted in a very general setup. However, we will make our findings transparent in the particular case, where the perturbation is a sequence of square integrable martingale differences whose conditional covariance converges to a random matrix Γ, almost surely, on $\mathbb{M}^{\mathrm{conv}}$. Here, we prove that under appropriate assumptions to be found in Theorem 2.6 the Cesàro average

$$\bar{X}_n = \frac{1}{n - n_0(n)} \sum_{k=n_0(n)+1}^{n} X_k$$

converges in the stable sense, on $\mathbb{M}^{\mathrm{conv}}$,

$$\sqrt{n}\,(\bar{X}_n - \bar{X}_n^*) \overset{\text{stably}}{\Longrightarrow} \left(Df(X_\infty)\big|_{N_{X_\infty}M}\right)^{-1} \Pi_{N_{X_\infty}M}\,\mathcal{N}(0,\Gamma),$$

where the right-hand side stands for the random distribution obtained when applying the orthogonal projection $\Pi_{N_{X_\infty}M}$ onto the normal space of M at X_∞ and the inverse of the restricted random mapping $Df(X_\infty)\big|_{N_{X_\infty}M} : N_{X_\infty}M \to N_{X_\infty}M$ (which will exist as consequence of a variant of the standard contractivity assumption) to a centered Gaussian random variable with covariance Γ. Note that the order of convergence is the same as for isolated attractors. Moreover, in the latter case the manifold M is zero dimensional and $N_{X_\infty}M = \mathbb{R}^d$ so that one recovers the classical result that, on $\mathbb{M}^{\mathrm{conv}}$,

$$\sqrt{n}\,(\bar{X}_n - X_\infty) \overset{\text{stably}}{\Longrightarrow} \left(Df(X_\infty)\right)^{-1} \mathcal{N}(0,\Gamma).$$

Here and in the main theorems, we use stable convergence *restricted to sets* that are not necessarily almost sure sets. The respective notion of convergence is introduced and analysed in detail in Section A.

Still, there is a crucial difference between the setting with isolated attractors and the one we discuss here. To explain this and later to do the proofs, we assume existence of particular local manifold representations $\Psi : U \to \mathbb{R}^d$ around some open sets $U \subset \mathbb{R}^d$ which allow us to associate every $x \in M \cap U$ with coordinates

$$\Psi(x) = \begin{pmatrix} \Psi_\zeta(x) \\ \Psi_\theta(x) \end{pmatrix} \in \mathbb{R}^{d_\zeta} \times \{0\}^{d_\theta} \subset \mathbb{R}^d$$

in such a way that for $x \in U$

$$\Psi(x^*) = \begin{pmatrix} \Psi_\zeta(x) \\ 0 \end{pmatrix}.$$

In the representation we thus have well separated directions. The tangential directions are the ones in $\mathbb{R}^{d_\zeta} \times \{0\}^{d_\theta}$ and the normal ones are the ones in $\{0\}^{d_\zeta} \times \mathbb{R}^{d_\theta}$ with $d_\theta = d - d_\zeta$. On the event that $(X_n)_{n \in \mathbb{N}_0}$ converges to some element of $U \cap M$ the sequence has all but finitely many entries in U. In the new coordinates the fluctuations in the normal direction will behave as in the classical theory whereas the fluctuations in the tangential direction are typically *larger* since there is no restoring force acting in this direction. This explains why we need to compare $\bar{X}_n$ with $\bar{X}_n^*$ and not X_∞ in the central limit theorem. However, since F is locally constant on M, this analysis is sufficient to derive a central limit theorem for the objective function value of the Ruppert-Polyak average $(F(\bar{X}_n))_{n \in \mathbb{N}}$. While the fluctuations in the tangential direction do not appear in the limit distribution, we will need to impose additional assumptions on the sequence of step-sizes to show that these effects are negligible. More explicitly, in the setting with the highest regularity (e.g. in the case where $F : \mathbb{R}^d \to \mathbb{R}$ is C^3 and M is a C^3-manifold) we allow step-sizes $\gamma_n = C_\gamma n^{-\gamma}$ with $C_\gamma > 0$ and $\gamma \in (\frac{3}{4}, 1)$. In the case of isolated attractors one typically allows exponents $\gamma \in (\frac{1}{2}, 1)$, see e.g. [35].

In the article, we use $\mathcal{O}$-notation. For a multivariate function (f_n) and a strictly positive function function (g_n) we write

$$f_n = \mathcal{O}(g_n) \quad \text{if and only if} \quad \sup_n \frac{|f_n|}{g_n} < \infty$$

and

$$f_n = o(g_n) \quad \text{if and only if} \quad \lim_{n \to \infty} \frac{|f_n|}{g_n} = 0$$

with the former notation making sense for arbitrary domains and the latter one for domains being subsets of $\mathbb{R}$. We also make use of the notation in a probabilistic sense, see Section B for details.

2 The central limit theorem

In this section, we introduce the main result of the article, a central limit theorem for the averaged Robbins-Monro scheme on $\mathbb{M}^{\text{conv}}$. We start with introducing the central definitions. Generally, for a C^1-submanifold $M \subset \mathbb{R}^d$ we denote by $T_x M$ the tangent space of M at $x \in M$ and by $N_x M = (T_x M)^\perp$ the normal space of M at x.

Definition 2.1. *A pair (F, M) consisting of a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ and a d_ζ-dimensional C^1-submanifold M of $\mathbb{R}^d$ is called* approximation problem *if the following holds*

(i) $DF\big|_M \equiv 0$

(ii) $f = DF$ *is continuously differentiable on* M *and*

(iii) for every $x \in M$*, the differential* $Df(x)$ *is symmetric and satisfies, for every* $v \in N_x M \backslash \{0\}$*,*

$$\langle v, Df(x)v \rangle < 0. \tag{2.1}$$

Set $d_\theta = d - d_\zeta$*.*

Remark 2.2. If (F, M) is an approximation problem, then, for every $x \in M$, the symmetric matrix $Df(x)$ admits an orthonormal basis of eigenvectors with the first d_ζ-vectors spanning the tangential space $T_x M$. By orthogonality, the remaining eigenvectors are in $N_x M$ so that the restricted mapping $Df(x)\big|_{N_x M}$ maps $N_x M$ into $N_x M$. As consequence of (2.1), the restricted mapping $Df(x)\big|_{N_x M} : N_x M \to N_x M$ is injective and, thus, one-to-one. See Proposition 34 and Proposition 35 of [16] for examples of approximation problems that arise in the training of neural networks.

Furthermore, we introduce a notion of regularity that entails error estimates for certain Taylor approximations in our proofs. We will express our assumptions on the vector field f and a certain local parametrisation of the manifold M in this notion.

Definition 2.3. *Let* $U \subset \mathbb{R}^d$ *be an open set,* $g : U \to \mathbb{R}^d$ *be a mapping and* $\alpha_g \in (0, 1]$*.*

(1) We say that g *has* regularity α_g *if* g *is continuously differentiable on* U *with* α_g*-Hölder continuous differential* Dg*.*

(2) Let, additionally, $M \subset \mathbb{R}^d$*. We say that* $g : U \to \mathbb{R}^d$ *has* regularity α_g around M *if*

(i) g *is continuously differentiable on* $M \cap U$ *with* α_g*-Hölder continuous differential and*

(ii) there exists a constant C *such that for all* $x \in M \cap U$ *and* $y \in U$

$$|g(y) - (g(x) + Dg(x)(y - x))| \leq C|y - x|^{1+\alpha_g}.$$

We introduce certain kind of parametrisations of the manifold that will appear in our proofs.

Definition 2.4. *Let* (F, M) *be an approximation problem and* $\alpha_f, \alpha_\Phi, \alpha_\Psi \in (0, 1]$*.*

(1) Let $U \subset \mathbb{R}^d$ *be an open set intersecting* M*. A* C^1*-diffeomorphism* $\Phi : U_\Phi \to U$ *is called* nice representation for M on U *if the following is true:*

(i) U_Φ *is a convex subset of* $\mathbb{R}^d$ *such that for* $(\zeta, \theta) \in \mathbb{R}^{d_\zeta} \times \mathbb{R}^{d_\theta}$

$$(\zeta, \theta) \in U_\Phi \;\Rightarrow\; (\zeta, 0) \in U_\Phi$$

and $\Phi(U_\Phi \cap (\mathbb{R}^{d_\zeta} \times \{0\}^{d_\theta})) = U \cap M$*.*

(ii) There exists a family $(P_x : x \in M \cap U)$ *of isometric isomorphisms* $P_x : \mathbb{R}^{d_\theta} \to N_x M$ *such that for every* $(\zeta, \theta) \in U_\Phi \subset \mathbb{R}^{d_\zeta} \times \mathbb{R}^{d_\theta}$

$$\Phi(\zeta, \theta) = \Phi(\zeta, 0) + P_{\Phi(\zeta,0)}(\theta). \tag{2.2}$$

(2) We say that (F, M) *has* regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$ *if for every* $x \in M$ *there exists a nice representation* $\Phi : U_\Phi \to U$ *of* M *on a neighbourhood* U *of* x *such that*

(i) the vector field $f\big|_U = DF\big|_U$ *has regularity* α_f *around* M*,*
(ii) the mapping Φ *has regularity* α_Φ *around* $\mathbb{R}^{d_\zeta} \times \{0\}^{d_\theta}$ *and*
(iii) its inverse $\Psi : U \to U_\Phi$ *has regularity* α_Ψ*.*

Further, an open set U *satisfying all the assumptions above are called nice representation for* M *on* U *with regularity* $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$*.*

It is natural to ask for simple criteria to decide whether an approximation problem has a certain regularity. We discuss this issue in the following remark.

Remark 2.5. 1. Let $\Psi : U \to V$ be a C^1-diffeomorphism with regularity $\alpha \in (0,1]$ and let $U' \subset \mathbb{R}^d$ be a bounded and connected open set with $\overline{U'} \subset U$. By Theorem 1.3.4 of [17], it follows that the inverse $\Psi^{-1}\big|_{\Psi(U')} : \Psi(U') \to U'$ has also regularity α. Hence, an approximation problem has regularity $(\alpha_f, \alpha, \alpha)$ if for every $x \in M$ there exists a nice representation $\Phi : U_\Phi \to U$ of M on a neighbourhood U of x such that

(a) the vector field $f\big|_U = DF\big|_U$ has regularity α_f around M,
(b') one of the mappings Φ or Ψ has regularity α.

2. Let (F, M) be an approximation problem, so that M is a C^3-manifold. Section C shows that for every $x \in M$ there exist a neighbourhood $U \subset \mathbb{R}^d$ of x and a nice representation $\Phi : U_\Phi \to U \in C^2$. Thus, after shrinking U_Φ we can guarantee that $D\Phi$ is Lipschitz and, again with Theorem 1.3.4 of [17], Φ is invertible with the differential of its inverse being a Lipschitz function. Hence, an approximation problem has regularity $(\alpha_f, 1, 1)$ if for every $x \in M$ there exists a neighbourhood U of x such that

(a) $f\big|_U$ has regularity α_f around M and
(b") M is a C^3-manifold.

Now we are able to state the main results.

Theorem 2.6. *Let (F, M) be an approximation problem and suppose that $(X_n)_{n\in\mathbb{N}_0}$ is the Robbins-Monro system and $(\bar{X}_n)$ the Ruppert-Polyak average as introduced in (1.1) and (1.2) with (D_n), (γ_n), (b_n), $(\bar{b}_n)$ and $(n_0(n))$ as in the introduction. Furthermore, let $\mathbb{M}^{\mathrm{conv}}$ denote the event that (X_n) converges to an element of M and denote by X_∞ its limit which is a well-defined and measurable function on $\mathbb{M}^{\mathrm{conv}}$. We consider the following assumptions:*

(A.1) Regularity. (F, M) has regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$, where $\alpha_f, \alpha_\Phi \in (0,1]$ and $\alpha_\Psi \in (\frac{1}{2}, 1]$.

(A.2) Assumptions on (γ_n) and (b_n). Set $\alpha = \alpha_\Psi \wedge \alpha_f \wedge \alpha_\Phi$ and $\alpha' = \alpha_\Psi \wedge \frac{1+\alpha}{2} > \frac{1}{2}$. Suppose that

$$\Big(1 - \frac{\alpha}{1+2\alpha}\Big) \vee \Big(1 - \frac{1}{2}\frac{\alpha_\Phi}{1+\alpha_\Phi}\Big) \vee \frac{1}{2\alpha'} < \gamma < 1 \ \text{ and } \ 1 + \rho > \gamma\alpha', \tag{2.3}$$

and set

$$\gamma_n = C_\gamma n^{-\gamma} \ \text{ and } \ b_n = n^\rho.$$

(A.3) Assumptions on $(n_0(n))$. $(n_0(n))_{n\in\mathbb{N}}$ is a $\mathbb{N}_0$-valued sequence with $0 \le n_0(n) < n$ for all $n \in \mathbb{N}$ that satisfies

$$n_0(n) = o(n) \ \text{ and } \ n_0(n)^{-1} = o\Big(n^{-\frac{1}{2\gamma-1}\frac{1}{1+\alpha_\Phi}} \wedge n^{-\frac{1}{\alpha}\frac{1-\gamma}{2\gamma-1}}\Big). \tag{2.4}$$

If $\rho < \gamma - 1$ we, additionally, assume that

$$n_0(n)^{-1} = o\Big(n^{-\frac{\frac{1}{1+\alpha_\Phi} - (1+\rho)}{\gamma-(1+\rho)}}\Big). \tag{2.5}$$

(A.4) Assumptions on D_n. For every $x \in M$ there exists an open neighbourhood $U \subset \mathbb{R}^d$ of x such that $(\mathbb{1}_U(X_{n-1})D_n)_{n\in\mathbb{N}}$ is a sequence of uniformly L^2-integrable martingale differences. Moreover,

$$\lim_{n\to\infty} \mathrm{cov}(D_n|\mathcal{F}_{n-1}) = \Gamma, \ \textit{almost surely, on } \mathbb{M}^{\mathrm{conv}}.$$

Under the above assumptions the following is true:

1. CLT for the coefficients. *On* $\mathbb{M}^{\mathrm{conv}}$*, one has*

$$\sqrt{n}\,(\bar{X}_n - \bar{X}_n^*) \overset{\text{stably}}{\Longrightarrow} \frac{\rho+1}{\sqrt{2\rho+1}} \big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1}\,\Pi_{N_{X_\infty}M}\,\mathcal{N}(0,\Gamma), \tag{2.6}$$

   where the right-hand side stands for the random distribution being obtained when applying the $\mathcal{F}_\infty$*-measurable linear transform* $\big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1}\,\Pi_{N_{X_\infty}M}$ *onto a normally distributed random variable* $\mathcal{N}(0,\Gamma)$ *with mean zero and covariance* Γ*.*

2. CLT for the F-performance. *On* $\mathbb{M}^{\mathrm{conv}}$*, one has*

$$2n(F(X_\infty) - F(\bar{X}_n)) \overset{\text{stably}}{\Longrightarrow} \Big|\frac{\rho+1}{\sqrt{2\rho+1}}\,\big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1/2}\,\Pi_{N_{X_\infty}M}\,\mathcal{N}(0,\Gamma)\Big|^2, \tag{2.7}$$

   where the right-hand side stands for the random distribution being obtained when applying the respective $\mathcal{F}_\infty$*-measurable operations onto a normally distributed random variable with mean zero and covariance* Γ*.*

If assumption (A.1) is true, there are feasible choices for γ *and* ρ *that satisfy (2.3) and for every such choice there exist feasible choices for* $(n_0(n))_{n\in\mathbb{N}}$ *satisfying (A.3).*

Theorem 2.6 is a special case of Theorem 2.9 below.

Remark 2.7. 1. It is straight-forward to verify that the factor $\frac{\rho+1}{\sqrt{2\rho+1}}$ appearing on the right-hand side of (2.6) and (2.7) is minimal for $\rho = 0$. Furthermore, irrespective of the choice of allowed parameters we always have $1 > \gamma\alpha'$ so that $\rho = 0$ is always a feasible choice, see (2.3). Thus, taking a *Cesàro average* is always optimal.

2. The choice of α's that leads to the least restrictions on the choice of γ are $\alpha_\Phi = 1$, $\alpha_\Psi = \frac{2}{3}$, $\alpha_f = \frac{1}{2}$. In that case all terms on the left-hand side of the γ-condition (2.3) equal $\frac{3}{4}$ so that we are allowed to choose γ in $(\frac{3}{4}, 1)$.

Remark 2.8. In this remark, we illustrate Theorem 2.6 in a particular optimisation problem. Let $F : \mathbb{R}^d \to \mathbb{R}$ be given in terms of the expectation

$$F(x) = \mathbb{E}[G(x, Y)],$$

where Y is a random variable taking values in a measurable space $\mathcal{Y}$ (we omit the σ-field in order to simplify notation) and $G : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is a product-measurable function satisfying the following regularity assumptions:

(i) F is C^3 with Lipschitz continuous differential f,

(ii) $\lim_{|x|\to\infty} F(x) = -\infty$,

(iii) for every $y \in \mathcal{Y}$ the function $G(\cdot, y)$ is C^1 and one has, for every $x \in \mathbb{R}^d$,

$$f(x) = \mathbb{E}[\nabla_x G(x, Y)],$$

(iv) the mapping

$$\mathcal{C} : \mathbb{R}^d \ni x \mapsto \mathrm{cov}(\nabla_x G(x, Y))$$

is continuous and

(v) there exist $q > 2$, $C_D < \infty$ and a locally bounded, measurable function $Q : \mathbb{R}^d \to (0, \infty)$ such that, for every $x \in \mathbb{R}^d$,

$$\mathbb{E}[|\nabla_x G(x, Y) - f(x)|^q]^{1/q} \le C_D\, Q(x).$$

To design a dynamical system $(X_n)_{n\in\mathbb{N}_0}$ as in (1.1) we fix $C_\gamma > 0$, $\gamma \in (\frac{3}{4}, 1)$, $m \in \mathbb{N}$ and an $\mathbb{R}^d$-valued random variable X_0, choose $(\gamma_n)_{n\in\mathbb{N}} = (C_\gamma n^{-\gamma})_{n\in\mathbb{N}}$ and an i.i.d. sequence $(Y_{n,i})_{n,i\in\mathbb{N}}$ of copies of Y that is also independent of X_0 and consider

$$D_n = \frac{1}{m_n}\sum_{i=1}^{m_n} \nabla_x G(X_{n-1}, Y_{n,i}) - f(X_{n-1}),$$

where $m_n = m \vee \lceil n^{-1/2} Q(X_{n-1})^2 \rceil$. Then $(D_n)_{n\in\mathbb{N}}$ defines a sequence of martingale differences w.r.t. the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ given by $\mathcal{F}_n = \sigma(X_0) \vee \sigma(Y_{k,i} : k, i \in \mathbb{N} \text{ with } k \le n)$ and $(X_n)_{n\in\mathbb{N}_0}$ satisfies

$$X_n = X_{n-1} + \gamma_n \frac{1}{m_n}\sum_{i=1}^{m_n} \nabla_x G(X_{n-1}, Y_{n,i}), \quad \text{for } n \in \mathbb{N}.$$

As consequence of the Burkholder-Davis-Gundy inequality there exists a universal constant $C_q > 0$ such that

$$\mathbb{E}[|D_n|^q|\mathcal{F}_{n-1}]^{1/q} \le C_q \frac{1}{m_n^{1/2}} \mathbb{E}[|\nabla_x G(x, Y) - f(x)|^q]^{1/q}\big|_{x=X_{n-1}} \le C_q C_D \frac{1}{m_n^{1/2}} Q(X_{n-1}).$$

Hence, for an arbitrary bounded, open set U, one has

$$\mathbb{E}[\mathbb{1}_U(X_{n-1})|D_n|^q]^{1/q} \le \frac{C_q C_D}{m^{1/2}} \sup_{x\in D} Q(x) < \infty$$

and $(1_U(X_{n-1})D_n)_{n\in\mathbb{N}}$ is uniformly L^2-integrable. Moreover, $\operatorname{cov}(D_n|\mathcal{F}_{n-1}) = \frac{1}{m_n}\mathcal{C}(X_{n-1})$ and, on the event $\{(X_n)_{n\in\mathbb{N}_0} \text{ converges}\}$, we have that $m_n \to m$ and, by continuity of $\mathcal{C}$, that

$$\operatorname{cov}(D_n|\mathcal{F}_{n-1}) = \frac{1}{m_n}\mathcal{C}(X_{n-1}) \overset{n\to\infty}{\longrightarrow} \frac{1}{m}\mathcal{C}(X_\infty) =: \Gamma.$$

This entails assumption (A.4) of Theorem 2.6. Let us verify the remaining assumptions.

Suppose that M is a C^3-manifold for which (F, M) is an approximation problem according to Definition 2.1. By Remark 2.5, (F, M) has regularity $(1, 1, 1)$. We choose $b_n \equiv 1$ and note that by Remark 2.7, (A.2) is satisfied. Since $1/(4\gamma - 2) < 1$ we can choose $(n_0(n))_{n\in\mathbb{N}}$ according to (A.3) and we consider the Césaro average

$$\bar{X}_n = \frac{1}{n - n_0(n)} \sum_{k=n_0(n)+1}^{n} X_k.$$

Consequently, we have, on $\mathbb{M}^{\mathrm{conv}}$,

$$2mn(F(X_\infty) - F(\bar{X}_n)) \overset{\text{stably}}{\Longrightarrow} \left|\left(Df(X_\infty)\big|_{N_{X_\infty}M}\right)^{-1/2} \Pi_{N_{X_\infty}M}\, \mathcal{N}(0, \mathcal{C}(X_\infty))\right|^2.$$

Let us discuss convergence of the dynamical system $(X_n)_{n\in\mathbb{N}_0}$.

(i) Locality of the process: First, we consider the event

$$\mathbb{L} = \Big\{\limsup_{n\to\infty} |X_n| < \infty\Big\}.$$

By choice of $(m_n)_{n\in\mathbb{N}}$ and property (v),

$$\mathbb{E}[|D_n|^2|\mathcal{F}_{n-1}]^{1/2} \le \frac{1}{\sqrt{m_n}} C_D\, Q(X_{n-1}) \le C_D\, n^{1/4}.$$

Verifying the assumptions of Lemma D.1 for the choice $(\sigma_n^{\mathrm{RM}})_{n\in\mathbb{N}} = (n^{\frac{1-2\gamma}{4}})_{n\in\mathbb{N}}$ we deduce that under assumptions (i) to (v), we have $\mathbb{P}(\mathbb{L}) = 1$. In particular, this implies

that, almost surely, $m_n = m$ for all but finitely many $n \in \mathbb{N}$. In the case that F satisfies (i) and (iii)-(v) but not necessarily property (ii), one can add an L^2-regularisation term, i.e., replace F by $\tilde{F}$ given by

$$\tilde{F}(x) = F(x) - \frac{a}{2}|x|^2,$$

for a fixed $a \in (0, \infty)$. Then, $D\tilde{F}(x) = \mathbb{E}[\nabla_x G(x, Y)] - ax$ is again a Lipschitz continuous function and the martingale noise remains the same. If $a > \|DF\|_{\mathrm{Lip}(\mathbb{R}^d)}$, then clearly, $\lim_{|x|\to\infty} F(x) = -\infty$. Note that this transformation typically affects the set of optimal points.

(ii) Convergence of the process: In [8] it is shown that $(X_n)_{n\in\mathbb{N}_0}$ almost surely converges, on $\mathbb{L}$, if every critical point of F satisfies locally a Łojasiewicz inequality, i.e. for all $x \in \{x' \in \mathbb{R}^d : f(x') = 0\}$ there exists a neighbourhood $U_x \subset \mathbb{R}^d$ of x, and parameters $\beta_x \in [\frac{1}{2}, 1)$, $\text{Ł}_x > 0$ such that, for all $x' \in U_x$, we have

$$|f(x')| \geq \text{Ł}_x |F(x) - F(x')|^{\beta_x}.$$

This assumption has the appeal that it is satisfied by every analytic function, see [26, 27]. Moreover, in Theorem 2.1 of [15] the following result is shown that resembles the situation of Theorem 2.6: Let $U \subset \mathbb{R}^d$ be an open set and assume that $F : U \to \mathbb{R}$ is C^2 and $M' = \{x \in U : f(x) = 0\}$ forms a $d_{\zeta'}$-dimensional manifold. If, for all $x \in M'$, we have $d_{\zeta'} = \dim(\ker(\mathrm{Hess}\, F(x)))$, then, for all $x \in M'$, x satisfies locally a Łojasiewicz inequality.

(iii) Non-convergence to unstable points: Let $M' \subset \mathbb{R}^d$ be a smooth manifold of critical points such that there exists a $C > 0$ with $\mathrm{Hess}\, F(x)$ has at least one positive eigenvalue, for all $x \in M'$, all negative eigenvalues are bounded from above by $-C$ and all positive eigenvalues are bounded from below by C. Then, if the martingale noise is uniformly bounded and uniformly exciting, meaning that there exists a $\tilde{C} > 0$ such that, for all $n \in \mathbb{N}$ and $u \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[\langle D_n, u\rangle^+ | \mathcal{F}_{n-1}] \geq \tilde{C}, \tag{2.8}$$

it holds that $\mathbb{P}(d(X_n, M') \to 0) = 0$, see Theorem 3 in [28]. If $(D_n)_{n\in\mathbb{N}}$ is not uniformly exciting then, for all $n \in \mathbb{N}$, one can consider $(\tilde{D}_n)_{n\in\mathbb{N}} = (D_n + \Lambda_n)_{n\in\mathbb{N}}$, where $(\Lambda_n)_{n\in\mathbb{N}}$ is a sequence of independent standard Gaussians, and set $(\tilde{\mathcal{F}}_n)_{n\in\mathbb{N}_0} = (\mathcal{F}_n \vee \sigma(\Lambda_1, \ldots, \Lambda_n))_{n\in\mathbb{N}_0}$. Then, $(\tilde{D}_n)_{n\in\mathbb{N}}$ satisfies (2.8) as well as the conditions necessary for proving locality and convergence.

We give a more general version of Theorem 2.6 which applies for a broad choice of step-sizes, averaging parameters and stochastic noises.

Theorem 2.9. *Let (F, M) be an approximation problem and suppose that $(X_n)_{n\in\mathbb{N}_0}$ is the Robbins-Monro system and $(\bar{X}_n)$ the Ruppert-Polyak average as introduced in (1.1) and (1.2) with (D_n), (γ_n), (b_n), $(\bar{b}_n)$ and $(n_0(n))$ as in the introduction. Let (σ_n^{RM}) and $(\delta_n^{\mathrm{diff}})$ be sequences of strictly positive reals and set*

$$\sigma_n = \frac{1}{\bar{b}_n} \sqrt{\sum_{l=n_0(n)+1}^{n} (b_l \delta_l^{\mathrm{diff}})^2}.$$

Furthermore, let $\mathbb{M}^{\mathrm{conv}}$ denote the event that (X_n) converges to an element of M and denote by X_∞ its limit which is a well-defined and measurable function on $\mathbb{M}^{\mathrm{conv}}$. We consider the following assumptions:

(B.1) Regularity. (F, M) has regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$, where $\alpha_f, \alpha_\Phi, \alpha_\Psi \in (0, 1]$.

(B.2) Technical assumptions on the parameters. *Suppose that* (γ_n) *is a monotonically decreasing sequence and*

$$n\gamma_n \to \infty, \quad \gamma_n \to 0,$$

$$\frac{b_{n+1}\gamma_n}{b_n\gamma_{n+1}} = 1 + o(\gamma_n), \quad \limsup_{n\to\infty} \frac{1}{\gamma_n} \frac{\sigma_{n-1}^{\mathrm{RM}} - \sigma_n^{\mathrm{RM}}}{\sigma_n^{\mathrm{RM}}} = 0, \quad \sigma_{n-1}^{\mathrm{RM}} \approx \sigma_n^{\mathrm{RM}}, \tag{2.9}$$

and for all sequences $(L(n))_{n\in\mathbb{N}}$ *with* $L(n) \leq n$ *and* $n - L(n) = o(n)$ *one has*

$$\lim_{n\to\infty} \frac{\sum_{k=L(n)+1}^{n} (b_k\delta_k^{\mathrm{diff}})^2}{\sum_{k=n_0(n)+1}^{n} (b_k\delta_k^{\mathrm{diff}})^2} = 0.$$

(B.3) Assumptions on $(n_0(n))$. $(n_0(n))_{n\in\mathbb{N}}$ *is a* $\mathbb{N}_0$*-valued sequence with* $0 \leq n_0(n) < n$ *for all* $n \in \mathbb{N}$ *that satisfies* $n_0(n) = o(n)$.

(B.4) Assumptions on D_n. *For every* $x \in M$*, there exist an open neighbourhood* $U \subset \mathbb{R}^d$ *of* x *so that* $(\mathbb{1}_U(X_{n-1})D_n)_{n\in\mathbb{N}}$ *is a sequence of square integrable, martingale differences satisfying for all* $\varepsilon > 0$*, on* $\mathbb{M}^{\mathrm{conv}}$*,*

$$\lim_{n\to\infty} (\delta_n^{\mathrm{diff}})^{-2} \mathrm{cov}(D_n|\mathcal{F}_{n-1}) = \Gamma, \text{ almost surely,}$$

$$\lim_{n\to\infty} (\sigma_n)^{-2} \sum_{m=n_0(n)+1}^{n} \frac{b_m^2}{\bar{b}_n^2} \mathbb{E}[\mathbb{1}_{\{|D_m|>\varepsilon\bar{b}_n\sigma_n/b_m\}}|D_m|^2|\mathcal{F}_{m-1}] = 0, \text{ in probability,} \tag{2.10}$$

and

$$\limsup_{n\to\infty} \Big(\frac{\sigma_n^{\mathrm{RM}}}{\sqrt{\gamma_n}}\Big)^{-1} \mathbb{E}[\mathbb{1}_U(X_{n-1})|D_n|^2]^{1/2} < \infty. \tag{2.11}$$

(B.5) Technical assumptions to control the error terms. *One has, as* $n \to \infty$*,*

$$\frac{b_{n_0(n)}}{\bar{b}_n\gamma_{n_0(n)}} \sigma_{n_0(n)}^{\mathrm{RM}} = o(\sigma_n), \tag{2.12}$$

$$(\varepsilon_n^{\mathrm{RM}})^{1+\alpha_\Phi} = o(\sigma_n), \tag{2.13}$$

for

$$\varepsilon_n^{\mathrm{RM}} := \sum_{k=n_0(n)+1}^{n} \big((\sqrt{\gamma_k}\sigma_k^{\mathrm{RM}})^{1+\alpha_\Psi} + \gamma_k(\sigma_{k-1}^{\mathrm{RM}})^{1+\alpha}\big) + \sqrt{\sum_{k=n_0(n)+1}^{n} \gamma_k(\sigma_k^{\mathrm{RM}})^2},$$

$$\varepsilon_n^{\mathrm{RP}} := \frac{1}{\bar{b}_n} \sum_{k=n_0(n)+1}^{n} b_k\big(\gamma_k^{-\frac{1-\alpha_\Psi}{2}}(\sigma_k^{\mathrm{RM}})^{1+\alpha_\Psi} + (\sigma_{k-1}^{\mathrm{RM}})^{1+\alpha} + \sigma_{k-1}^{\mathrm{RM}}(\varepsilon_n^{\mathrm{RM}})^{\alpha}\big) = o(\sigma_n) \tag{2.14}$$

and

$$\Big(\frac{1}{\bar{b}_n} \sum_{m=n_0(n)+1}^{n} b_m(\sigma_m^{\mathrm{RM}})^2\Big)^{(1+\alpha_\Phi)/2} = o(\sigma_n). \tag{2.15}$$

Under the above assumptions the following is true:

1. CLT for the coefficients. *On* $\mathbb{M}^{\mathrm{conv}}$*, one has*

$$\sigma_n^{-1}\,(\bar{X}_n - \bar{X}_n^*) \overset{\text{stably}}{\Longrightarrow} \left(Df(X_\infty)\big|_{N_{X_\infty}M}\right)^{-1} \Pi_{N_{X_\infty}M}\,\mathcal{N}(0,\Gamma), \tag{2.16}$$

where the right-hand side stands for the random distribution being obtained when applying the $\mathcal{F}_\infty$*-measurable transform* $\left(Df(X_\infty)\big|_{N_{X_\infty}M}\right)^{-1} \Pi_{N_{X_\infty}M}$ *onto a normally distributed random variable with mean zero and covariance* Γ*.*

2. CLT for the F-performance. *On* $\mathbb{M}^{\mathrm{conv}}$*, one has*

$$2\sigma_n^{-2}(F(X_\infty) - F(\bar{X}_n)) \overset{\text{stably}}{\Longrightarrow} \left|\left(Df(X_\infty)\big|_{N_{X_\infty}M}\right)^{-1/2} \Pi_{N_{X_\infty}M}\,\mathcal{N}(0,\Gamma)\right|^2, \tag{2.17}$$

where the right-hand side stands for the random distribution being obtained when applying the respective $\mathcal{F}_\infty$*-measurable operations onto a normally distributed random variable with mean zero and covariance* Γ*.*

Remark 2.10. If we, additionally, assume in the theorem that there exists $L > 0$, so that for every $x \in M$, the differential $Df(x)$ satisfies, for every $v \in N_xM$,

$$\langle v, Df(x), v\rangle \leq -L|v|^2,$$

then assumption (2.9) can be relaxed to

$$\frac{b_{n+1}\gamma_n}{b_n\gamma_{n+1}} = 1 + o(\gamma_n),\ \ \limsup_{n\to\infty} \frac{1}{\gamma_n}\,\frac{\sigma_{n-1}^{\mathrm{RM}} - \sigma_n^{\mathrm{RM}}}{\sigma_n^{\mathrm{RM}}} < L,\ \ \sigma_{n-1}^{\mathrm{RM}} \approx \sigma_n^{\mathrm{RM}}.$$

We outline the structure of the proof of Theorem 2.9. Overall, we follow the martingale CLT approach introduced in [36]. The proof is based on a martingale CLT given in [19] that is generalised in the appendix (see Theorem A.5) to stable convergence restricted to non-trivial events. To prove the result we first analyse linear systems (the particular case where f is a matrix-multiplication) in Section 5. The general results are proved by representing the iterates of the Robbins-Monro scheme in an appropriate coordinate system and comparing the non-linear system with an appropriate linear system. Appropriate coordinate representations are introduced and analysed in Section 3. In order to control the perturbations, we derive an L^2-estimate in Section 4, see Theorem 4.1. The proof of the main results is carried out in Section 7, where the representation of the orthogonal coordinates in terms of a perturbed system can be found in (7.1). In order to keep the presentation simple, we collect and prove further technical estimates in Section 6.

3 Geometric preliminaries

In this section, we discuss some geometric properties of the d_ζ-dimensional stable manifold M. First, we derive that for an approximation problem (F, M) in sufficiently small neighbourhoods of M the strength of attraction is uniformly bounded away from zero. Afterwards, we discuss the well-definedness and regularity of the projection that maps every point to its nearest neighbour in M.

Definition 3.1. *Let* (F, M) *be an approximation problem. We call an open and bounded set* $U \subset \mathbb{R}^d$ *intersecting* M (F, M)-attractor *with stability* L *and bound* C*, for* $C \geq L > 0$*, if*

(i) $\bar{M} \cap U = M \cap U$ *and*

(ii) for every $x \in M \cap U$ *and* $v \in N_xM$

$$-C|v|^2 \leq \langle v, Df(x)v\rangle \leq -L|v|^2. \tag{3.1}$$

Lemma 3.2. *Let (F, M) be an approximation problem and $x \in M$, then x admits an open neighbourhood U and constants $C, L > 0$ such that U is an (F, M)-attractor with stability L and bound C.*

Proof. Let $\Psi : U \to U_\Phi$ be a C^1-diffeomorphism with U being an open neighbourhood of x and $U_\Phi \subset \mathbb{R}^d$ such that $\Psi(U \cap M) = U_\Phi \cap (\mathbb{R}^{d_\zeta} \times \{0\}^{d_\theta})$.

First, we show that $\bar{M} \cap U = M \cap U$. Let $z \in \bar{M} \cap U$. Then there exists a $M \cap U$-valued sequence $(z_n)_{n\in\mathbb{N}}$ with $z_n \to z$. Thus,

$$\Psi(z_n) = \begin{pmatrix} \Psi_\zeta(z_n) \\ 0 \end{pmatrix} \to \Psi(z) \quad \text{with} \quad \Psi_\zeta(x_n) = (\Psi_1(x_n), \dots, \Psi_{d_\zeta}(x_n)).$$

Consequently, $\Psi_i(z) = 0$ for all $i > d_\zeta$ and, hence, $z \in M$.

Second we show that for every bounded set $U' \subset U$ with $\overline{U'} \subset U$ there exist $C, L > 0$ such that for all $z \in M \cap \overline{U'}$ and $v \in N_x M$

$$-C|v|^2 \le \langle v, Df(z)v\rangle \le -L|v|^2.$$

It suffices to show that

$$\mathcal{C} := \{(z, v) \in \mathbb{R}^d \times \mathbb{R}^d : z \in M \cap \overline{U'}, v \in N_z M, |v| = 1\}$$

is a compact set since then C and L can be chosen as

$$-C = \min_{(z,v)\in\mathcal{C}} \langle v, Df(z)v\rangle \text{ and } -L = \max_{(z,v)\in\mathcal{C}} \langle v, Df(z)v\rangle$$

with the minimum and maximum both being obtained and being in $(-\infty, 0)$. Since $\mathcal{C}$ is bounded it remains to prove closedness. Let $(z_n, v_n)_{n\in\mathbb{N}}$ be a $\mathcal{C}$-valued sequence that converges to (z, v). Since $M \cap \overline{U'} = \overline{M} \cap \overline{U'}$ is compact we have that $z \in M \cap \overline{U'}$. We denote by Φ the inverse of Ψ and note that for all vectors $w \in \mathbb{R}^{d_\zeta} \times \{0\}$, $\partial_w \Phi(z_n)$ is in $T_{z_n} M$ which is perpendicular to $v_n \in N_{z_n} M$. Hence,

$$0 = \langle \partial_w \Phi(z_n), v_n\rangle \to \langle \partial_w \Phi(z), v\rangle$$

and $v \perp \partial_w \Phi(z)$. Since the considered vectors $\partial_w \Phi(z)$ span the tangent space $T_z M$ it follows that $v \in (T_z M)^\perp = N_z M$ and we are done. □

Remark 3.3. Let (F, M) be an approximation problem. Then, for $x \in M$, equation (3.1) is satisfied for all $v \in N_x M$ if the spectrum of $Df(x)$ restricted to $N_x M$ is contained in $[-C, -L]$. Indeed, there is always an orthonormal basis of eigenvectors $v_1, \dots, v_d$ with $v_1, \dots, v_{d_\zeta}$ spanning $T_x M$ and $v_{d_\zeta+1}, \dots, v_d$ spanning $N_x M$ and the equivalence follows by elementary linear algebra.

The remark entails the following corollary.

Corollary 3.4. *Let U be a (F, M)-attractor with stability L and bound C, $x \in U \cap M$ and $v \in N_x M$. Then for every $\gamma \in [0, C^{-1}]$ one has*

$$|v + \gamma Df(x)v| \le (1 - \gamma L)|v|.$$

Proof. By Remark 3.3, the spectrum of the restricted mapping $Df(x)\big|_{N_x M} : N_x M \to N_x M$ is contained in $[-C, -L]$. Hence, the spectrum of the restricted mapping $(\mathrm{id} + \gamma Df(x))\big|_{N_x M}$ is contained in $[1 - \gamma C, 1 - \gamma L] \subset [0, 1 - \gamma L]$ which immediately implies the result since the latter mapping is diagonalizable. □

For the next proposition we need the additional assumption, that the error of the first-order Taylor expansion of f is locally uniform. If f has regularity α_f around M for some $\alpha_f \in (0, 1]$, this follows immediately.

Proposition 3.5. *Let $U \subset \mathbb{R}^d$ be an (F, M)-attractor with stability L and bound C. Suppose that for $x \in U$ and $x' \in U \cap M$*

$$f(x) = Df(x')(x - x') + o(|x - x'|) \text{ as } |x - x'| \to 0$$

with the small o term being uniform in the choice of x and x'. Then for every $L' \in (0, L)$ and $\delta > 0$ there exists $\rho > 0$ such that for

$$U_\delta^\rho := \bigcup_{\substack{y \in M: \\ d(y, U^c) > \delta}} B_\rho(y) \tag{3.2}$$

one has for all $x \in U_\delta^\rho$ and $\gamma \in [0, C^{-1}]$

$$d(x + \gamma f(x), M) \leq (1 - \gamma L')d(x, M). \tag{3.3}$$

Proof. Choose $\rho \in (0, \frac{1}{2}\delta]$ such that for all $x, x' \in U$ with $x' \in M$ and $|x' - x| \leq \rho$

$$|f(x) - Df(x')(x - x')| \leq (L - L')|x - x'|.$$

Let $x \in U_\delta^\rho$. Then, by definition of U_δ^ρ there exists $x' \in M$ with $d(x, x') < \rho$ and $d(x', U^c) > \delta$. We denote by $z \in \bar{M}$ an element with

$$d(x, z) = d(x, \bar{M}) = d(x, M) < \rho.$$

Note that $d(x', z) \leq d(x', x) + d(x, z) < 2\rho \leq \delta$ so that $z \in B_\delta(x') \subset U$ and, hence, $z \in \bar{M} \cap U = M \cap U$. Take $v \in T_z M$ and a C^1-curve $\gamma : (-1, 1) \to M$ with $\gamma(0) = z$ and $\dot{\gamma}(0) = v$. Then, since $t \mapsto d(\gamma(t), x)^2$ has a minimum in 0 we get that

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\, d(\gamma(t), x)^2 \Big|_{t=0} = 2\langle z - x, v \rangle.$$

Thus $x - z \in N_z M$. With Lemma 3.4 we obtain that for $\gamma \in [0, C^{-1}]$

$$\begin{aligned} d(x + \gamma f(x), M) &\leq d(x + \gamma f(x), z) \leq |(\mathrm{Id} + \gamma Df(z))(x - z)| + \gamma |f(x) - Df(z)(x - z)| \\ &\leq (1 - \gamma L)|x - z| + \gamma(L - L')|x - z| = (1 - \gamma L')d(x, M). \end{aligned}$$

□

We consider the projection onto M which is defined as follows. For $x \in \mathbb{R}^d$ we set

$$x^* = \mathrm{argmin}_{y \in M} d(x, y),$$

if there is a unique minimizer.

We will show that for a nice representation $\Phi : U_\Phi \to U$ of M on some open and bounded set U (in the sense of Definition 2.4) and its inverse Ψ we have

$$x^* = \Phi(\Psi_\zeta(x), 0)$$

for all $x \in U$ that are sufficiently close to M. Here, Ψ_ζ represents the first d_ζ coordinates of Ψ, that is $\Psi_\zeta(x) = (\Psi_1(x), \ldots, \Psi_{d_\zeta}(x))$ for $x \in U$.

Lemma 3.6. *Let $\delta > 0$ and $U \subset \mathbb{R}^d$ an open and bounded set and $\Phi : U_\Phi \to U$ a nice representation for M on U.*

(i) There exists $\rho \in (0, \delta/4]$ such that for every $x \in M$ with $d(x, U^c) > \delta/2$ and $\theta \in \mathbb{R}^{d_\theta}$ with $|\theta| < \rho$ it holds

$$\Psi(x) + \begin{pmatrix} 0 \\ \theta \end{pmatrix} = \begin{pmatrix} \Psi_\zeta(x) \\ \theta \end{pmatrix} \in U_\Phi. \tag{3.4}$$

(ii) Suppose that $\rho > 0$ is as in (i). Then, for every $x \in U_\delta^\rho$, x^* is well-defined and one has the following:

(a) $x^* = \Phi(\Psi_\zeta(x), 0)$,
(b) the segment connecting x and x^* lies in U and
(c) $|\Psi_\theta(x)| = d(x, M)$ and $d(x^*, U^c) > \delta/2$.

Proof. **(i):** Let $\delta > 0$ and note that

$$M' := \{x \in \bar{M} : d(x, U^c) \geq \delta/2\} \subset \bar{M} \cap U = M \cap U$$

is a compact set. Hence, the continuous mapping

$$M' \ni x \mapsto d(\Psi(x), U_\Phi^c)$$

attains its minimum, say ρ', which is strictly positive since Ψ does not attain values in the closed set U_Φ^c. Obviously, property (i) holds for $\rho = \min(\rho', \delta/4)$.

(ii): Let $\rho \in (0, \delta/4]$ as in (i) and let $x \in U_\delta^\rho$. First we show that an element $z \in \bar{M}$ with

$$d(x, z) = d(x, \bar{M}) = d(x, M)$$

lies in $M \cap U$ and satisfies $x - z \in N_z M$. By definition of U_δ^ρ there exists $x' \in M$ with $d(x, x') < \rho$ and $d(x', U^c) > \delta$. Thus $d(x', z) \leq d(x', x) + d(x, z) < 2\rho$ and $d(z, U^c) \geq d(x', U^c) - d(x', z) > \delta - 2\rho \geq \delta/2$ so that $z \in U$ and hence also $z \in \bar{M} \cap U = M \cap U$.

Take $v \in T_z M$ and a C^1-curve $\gamma : (-1, 1) \to M \cap U$ with $\gamma(0) = z$ and $\dot{\gamma}(0) = v$. Then, since $t \mapsto d(\gamma(t), x)^2$ has a minimum in 0 we get that

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\, d(\gamma(t), x)^2 \Big|_{t=0} = 2\langle z - x, v \rangle.$$

Thus $x - z \in N_z M$. We recall that $|x - z| = d(x, z) < \rho$ so that as consequence of the representation property (2.2) there exists $\theta \in \mathbb{R}^{d_\theta}$ with $|\theta| = |x - z| < \rho$ and

$$x = z + P_z(\theta).$$

Moreover, recalling that $d(z, U^c) > \delta/2$ we get with (i) that $(\Psi_\zeta(z), \theta)$ is in U_Φ and hence $x = \Phi(\Psi_\zeta(z), \theta)$. An application of Ψ_ζ yields that $\Psi_\zeta(x) = \Psi_\zeta(z)$ so that

$$z = \Phi(\Psi_\zeta(z), 0) = \Phi(\Psi_\zeta(x), 0)$$

is the unique minimizer and $x^* = z$. Furthermore, with (3.4) the segment connecting x^* and x, which is $\gamma : [0, 1] \to \mathbb{R}^d$, $t \mapsto \Phi(\Psi_\zeta(x), t\theta)$ lies in U and

$$|\Psi_\theta(x)| = |\theta| = d(z, x) = d(x, M). \qquad \square$$

Proposition 3.7. *Let (F, M) be an approximation problem with regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$ and suppose that all assumptions of Theorem 2.9 are satisfied. We call a triple (U, δ, ρ) consisting of an open set $U \subset \mathbb{R}^d$ and $\delta, \rho > 0$* feasible*, if*

- *there exists a nice representation $\Phi : U_\Phi \to U$ for M on U with regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$,*
- *U is an (F, M)-attractor with stability L and bound C for some values $L, C > 0$,*
- *$(\mathbb{1}_U(X_{n-1})D_n)_{n \in \mathbb{N}}$ is a sequence of L^2-martingale differences satisfying (2.11),*
- *$\delta > 0$ and $\rho \in (0, \delta/4]$ are such that (i) of Lemma 3.6 is true and inequality (3.3) holds for a $L' \in (0, L)$.*

Then there exists a countable set of feasible triples (U, δ, ρ) *such that the respective subsets* U_δ^ρ *of* $\mathbb{R}^d$ *cover the manifold* M*.*

Proof. For every $x \in \mathbb{R}^d$ and every feasible triple (U, δ, ρ) we denote by

$$R_x(U, \delta, \rho) = \sup\{r \geq 0 : B_x(r) \subset U_\delta^\rho\}$$

the radius of the triple (U, δ, ρ) at x. Note that by definition for $x, y \in \mathbb{R}^d$, $|R_x(U, \delta, \rho) - R_y(U, \delta, \rho)| \leq |x - y|$ so that the function

$$\mathbb{R}^d \ni x \mapsto R_x = \sup\{R_x(U, \delta, \rho) : (U, \delta, \rho) \text{ is feasible}\}$$

is Lipschitz continuous with Lipschitz constant 1. (Possibly, all function values are infinite.)

Now fix a $\kappa > 0$ and a countable set $\mathbb{I}_\kappa \subset \mathbb{R}^d$ such that

$$\bigcup_{z \in \mathbb{I}_\kappa} B_z(\kappa/3) = \mathbb{R}^d.$$

We construct a collection $\mathcal{U}_\kappa$ of feasible triples as follows. For every $z \in \mathbb{I}_\kappa$ with $R_z \geq 2\kappa/3$ we add a triple with z-radius greater or equal to $\kappa/2$. For every $z \in \mathbb{I}_\kappa$ with $R_z < 2\kappa/3$ we do not add a triple. Then $\mathcal{U}_\kappa$ is countable and for every $x \in M$ with $R_x \geq \kappa$ there exists a $z \in \mathbb{I}_\kappa$ with $|x - z| \leq \kappa/3$. Hence $R_z \geq 2\kappa/3$ and we thus added a triple (U, δ, ρ) with z-radius greater or equal to $\kappa/2$ which obviously contains x. Consequently, $\mathcal{U}_\kappa$ is a countable set of feasible triples that covers at least $\{x \in M : R_x \geq \kappa\}$. By a diagonalisation argument, we obtain a countable set $\bigcup_{n \in \mathbb{N}} \mathcal{U}_{1/n}$ of feasible triples that covers $\bigcup_{n \in \mathbb{N}} \{x \in M : R_x \geq 1/n\} = M$. □

Remark 3.8. We consider the setting of Theorem 2.9. Let $\mathcal{U}$ be a countable set of feasible triples that covers M as in Lemma 3.7. For a feasible triple (U, δ, ρ) the set U_δ^ρ is open and we consider the event $\mathbb{U}_{\delta,\rho}^{\text{conv}}$ that (X_n) converges to an element of $M \cap U_\delta^\rho$. Then the covering property of $\mathcal{U}$ ensures that

$$\mathbb{M}^{\text{conv}} = \bigcup_{(U,\delta,\rho) \in \mathcal{U}} \mathbb{U}_{\delta,\rho}^{\text{conv}}$$

and by Lemma A.3 the proof of Theorem 2.9 is achieved once we showed stable convergence on $\mathbb{U}_{\delta,\rho}^{\text{conv}}$ for general feasible triples (U, δ, ρ).

4 L^2-error bounds

In this chapter, we control the behaviour of the Robbins-Monro scheme around an (F, M)-attractor at late times in terms of the distance to M in the L^2-norm. We will later need these estimates to control errors that we infer when comparing the original dynamical system with a linearised one.

As in the chapters before, let (F, M) be an approximation problem and let $U \subset \mathbb{R}^d$ be an (F, M)-attractor with stability L and bound C. We denote by $f = DF$ the Jacobi matrix of F and consider a dynamical system (X_n) given by

$$X_n = X_{n-1} + \gamma_n(f(X_{n-1}) + \underbrace{R_n + D_n}_{=U_n}) \tag{4.1}$$

with

- $X_0 \in \mathbb{R}^d$ is a fixed deterministic starting value,
- R_n being $\mathcal{F}_{n-1}$-measurable and

- D_n is $\mathcal{F}_n$-measurable and $(\mathbb{1}_U(X_{n-1})D_n)_{n\in\mathbb{N}}$ is a sequence of square integrable martingale differences.

Thus, in this chapter we also allow the process to have a predictable bias which should be of lower order than the martingale noise. This assumption will be made precise in the following theorem. We obtain the process introduced in (1.1) by choosing $R_n \equiv 0$.

Theorem 4.1. *Let $U \subset \mathbb{R}^d$ be an (F, M)-attractor with stability L and bound C. Suppose that for $x \in U$ and $x' \in U \cap M$*

$$f(x) = Df(x')(x - x') + o(|x - x'|) \text{ as } |x - x'| \to 0$$

with the o-term being uniform in x and x'. Let $(\gamma_n)_{n\in\mathbb{N}}$ and $(\sigma_n)_{n\in\mathbb{N}}$ sequences of strictly positive reals with $\lim_{n\to\infty} \gamma_n = 0$, $\sum_{n=1}^{\infty} \gamma_n = \infty$ and

$$L'' := \limsup_{n\to\infty} \frac{1}{\gamma_n} \frac{\sigma_{n-1} - \sigma_n}{\sigma_n} < L \tag{4.2}$$

and suppose that $(X_n)_{n\in\mathbb{N}_0}$ satisfies recursion (4.1). Let $\delta, \rho > 0$ be such that Prop. 3.5 is true for a $L' \in (L'', L)$, that is for all $x \in U_\delta^\rho$ and $\gamma \in [0, C^{-1}]$ one has

$$d(x + \gamma f(x), M) \le (1 - \gamma L')d(x, M).$$

Furthermore, assume that

$$\limsup_{n\to\infty} \sigma_n^{-1} \mathbb{E}\big[\mathbb{1}\{X_{n-1} \in U_\delta^\rho\}|R_n|^2\big]^{1/2} < \infty \tag{4.3}$$

and

$$\limsup_{n\to\infty} \Big(\frac{\sigma_n}{\sqrt{\gamma_n}}\Big)^{-1} \mathbb{E}[\mathbb{1}\{X_{n-1} \in U_\delta^\rho\}|D_n|^2]^{1/2} < \infty. \tag{4.4}$$

Then, there exists a $\tilde{C} \in (0, \infty)$ such that for all $N \in \mathbb{N}$,

$$\limsup_{n\to\infty} \sigma_n^{-1} \mathbb{E}\big[\mathbb{1}_{\{X_m \in U_\delta^\rho \text{ for } m=N,\dots,n-1\}} d(X_n, M)^2\big]^{1/2} < \tilde{C}. \tag{4.5}$$

Proof. Let $L' \in (L'', L)$ and $\delta, \rho > 0$ as in the theorem. By monotonicity it suffices to restrict attention to large N. For sufficiently large constants C_1 and C_2 we can fix $N_0 \in \mathbb{N}$ such that for all $n \ge N_0$

$$\gamma_n \le C^{-1},\ \mathbb{E}\big[\mathbb{1}\{X_{n-1} \in U_\delta^\rho\}|R_n|^2\big] \le C_1\sigma_n^2 \text{ and } \mathbb{E}[\mathbb{1}\{X_{n-1} \in U_\delta^\rho\}|D_n|^2] \le C_2\frac{\sigma_n^2}{\gamma_n}. \tag{4.6}$$

Now fix $N \ge N_0$ and consider

$$\mathbb{U}_n = \{\forall l = N, \dots, n : X_l \in U_\delta^\rho\}, \text{ for } n \ge N.$$

One has for $n > N$

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_{\mathbb{U}_{n-1}} d(X_{n-1} + \gamma_n(f(X_{n-1}) + R_n + D_n), M)^2] \\
&\quad \le \underbrace{\mathbb{E}[\mathbb{1}_{\mathbb{U}_{n-1}} d(X_{n-1} + \gamma_n(f(X_{n-1}) + R_n), M)^2]}_{=:I_1(n)} + \underbrace{\gamma_n^2\, \mathbb{E}[\mathbb{1}_{\mathbb{U}_{n-1}}|D_n|^2]}_{=:I_2(n)}.
\end{aligned} \tag{4.7}$$

Moreover, by (3.3) one has on the event $\mathbb{U}_{n-1}$ for arbitrary $a > 0$

$$\begin{aligned}
&d(X_{n-1} + \gamma_n(f(X_{n-1}) + R_n), M)^2 \\
&\quad \le (1 - L'\gamma_n)^2 d(X_{n-1}, M)^2 + 2\gamma_n d(X_{n-1}, M)|R_n| + \gamma_n^2|R_n|^2 \\
&\quad \le ((1 - L'\gamma_n)^2 + a\gamma_n)\, d(X_{n-1}, M)^2 + (\frac{1}{a}\gamma_n + \gamma_n^2)|R_n|^2.
\end{aligned}$$

Consequently, with (4.6)

$$I_1(n) \leq ((1-L'\gamma_n)^2 + a\gamma_n)\,\mathbb{E}[\mathbb{1}_{\mathbb{U}_{n-1}} d(X_{n-1}, M)^2] + C_1(\frac{1}{a}\gamma_n + \gamma_n^2)\sigma_n^2.$$

Now note that as $n \to \infty$, $(1-L'\gamma_n)^2 = 1 - 2L'\gamma_n + o(\gamma_n)$. Moreover,

$$\frac{\sigma_{n-1}}{\sigma_n} = 1 + \frac{\sigma_{n-1} - \sigma_n}{\sigma_n} \leq 1 + L''\gamma_n + o(\gamma_n)$$

so that

$$\begin{aligned}\frac{\sigma_{n-1}^2}{\sigma_n^2}((1-L'\gamma_n)^2 + a\gamma_n) &\leq (1 + 2L''\gamma_n + o(\gamma_n))(1 - (2L' - a)\gamma_n + o(\gamma_n)) \\ &= 1 - (2L' - 2L'' - a)\gamma_n + o(\gamma_n).\end{aligned}$$

Recall that $L' > L''$ and we fix $a, b > 0$ such that $2L' - 2L'' - a > b$. Then, for sufficiently large $n \in \mathbb{N}$

$$\frac{\sigma_{n-1}^2}{\sigma_n^2}((1-L'\gamma_n)^2 + a\gamma_n) \leq 1 - b\gamma_n$$

and by increasing N we can guarantee that the previous inequality holds for all $n > N$. Thus,

$$\sigma_n^{-2} I_1(n) \leq (1-b\gamma_n)\sigma_{n-1}^{-2}\mathbb{E}[\mathbb{1}_{\mathbb{U}_{n-1}} d(X_{n-1}, M)^2] + C_1(\frac{1}{a} + \frac{1}{C})\gamma_n.$$

Additionally, we get with (4.6) that $\sigma_n^{-2} I_2(n) \leq C_2\gamma_n$. This implies that the expectation

$$\varphi_n := \sigma_n^{-2}\mathbb{E}[\mathbb{1}_{\mathbb{U}_{n-1}} d(X_n, M)^2] \qquad (n \geq N)$$

satisfies for $n > N$

$$\varphi_n \leq \sigma_n^{-2}(I_1(n) + I_2(n)) \leq (1-b\gamma_n)\varphi_{n-1} + \underbrace{(C_1(a^{-1} + C^{-1}) + C_2)}_{=:C_3}\gamma_n.$$

It follows that

$$\varphi_n - \frac{C_3}{b} \leq (1-b\gamma_n)\Big(\varphi_{n-1} - \frac{C_3}{b}\Big)$$

and by iteration that

$$\varphi_n - \frac{C_3}{b} \leq \Big(\varphi_N - \frac{C_3}{b}\Big)\prod_{l=N+1}^{n}(1-b\gamma_l) \to 0,$$

where convergence follows since $\sum_{l=N+1}^{\infty}\gamma_l = \infty$. Therefore,

$$\limsup_{n\to\infty}\varphi_n \leq \frac{C_3}{b}.$$

Note that the statement remains valid with the same constant on the right-hand side when increasing N. □

5 The Ruppert-Polyak system for linear systems

In this section, we provide a central limit theorem for a particular linear system. It will be the main technical tool for proving Theorem 2.9. More explicitly, we will show that on the level of coordinate mappings the system is approximated up to lower terms by the system analysed here.

Again $(\gamma_n)_{n\in\mathbb{N}}$ denotes a monotonically decreasing sequence of non-negative reals which converges to 0. Additionally, $(n_0(n))_{n\in\mathbb{N}}$ is an increasing $\mathbb{N}_0$-valued sequence with $n_0(n) \leq n$ that tends to infinity and for each $n \in \mathbb{N}$, let H_n be a $\mathcal{F}_{n_0(n)}$-measurable matrix. We set for $n, i, j \in \mathbb{N}$ with $i \leq j$

$$\mathcal{H}_n[i,j] = \prod_{r=i+1}^{j} (\mathbb{1} + \gamma_r H_n) \text{ and } \bar{\mathcal{H}}_n[i,j] = \sum_{r=i}^{j} \frac{\gamma_i b_r}{b_i} \mathcal{H}_n[i,r]. \tag{5.1}$$

Based on a sequence $(\mathcal{D}_l)_{l\in\mathbb{N}}$ of $\mathbb{R}^d$-valued random variables we consider the dynamical system $(\Xi_n)_{n\in\mathbb{N}}$ with

$$\Xi_n := \frac{1}{\bar{b}_n} \sum_{i=n_0(n)+1}^{n} b_i \bar{\mathcal{H}}_n[i,n]\, \mathcal{D}_i \tag{5.2}$$

and

$$\bar{b}_n = \sum_{i=n_0(n)+1}^{n} b_i.$$

Theorem 5.1. *Let $A \in \mathcal{F}_\infty$ and $(\delta_n)_{n\in\mathbb{N}}$ be a sequence of strictly positive reals. We assume the following assumptions:*

1. Technical assumptions on the parameters.

   $$n\gamma_n \to \infty, \quad \frac{b_{n+1}\gamma_n}{b_n\gamma_{n+1}} = 1 + o(\gamma_n)$$

   and for all sequences $(L(n))_{n\in\mathbb{N}}$ with $n_0(n) \leq L(n) \leq n$ and $n - L(n) = o(n)$ one has

   $$\lim_{n\to\infty} \frac{\sum_{k=L(n)+1}^{n} (b_k\delta_k)^2}{\sum_{k=n_0(n)+1}^{n} (b_k\delta_k)^2} = 0.$$

2. Assumptions on H_n. *$(H_n)_{n\in\mathbb{N}}$ is a sequence of symmetric matrices with each H_n being $\mathcal{F}_{n_0(n)}$-measurable and*

   $$\lim_{n\to\infty} H_n = H, \text{ almost surely, on } A,$$

   for a random symmetric matrix H with $\max \sigma(H) < 0$.

3. Assumptions on $\mathcal{D}_k$. *$(\mathcal{D}_k)_{k\in\mathbb{N}}$ is a sequence of square integrable martingale differences that satisfies for a random matrix Γ, on A,*

   (a) $\lim_{m\to\infty} \left\| \operatorname{cov}(\delta_m^{-1}\mathcal{D}_m | \mathcal{F}_{m-1}) - \Gamma \right\| = 0$, *almost surely, and*

   (b) for $\sigma_n = \frac{1}{\bar{b}_n} \sqrt{\sum_{m=n_0(n)+1}^{n} (b_m\delta_m)^2}$ *and all $\varepsilon > 0$, one has*

   $$\lim_{n\to\infty} \sigma_n^{-2} \sum_{m=n_0(n)+1}^{n} \frac{b_m^2}{\bar{b}_n^2} \mathbb{E}\big[\mathbb{1}_{\{|\mathcal{D}_m| > \frac{\varepsilon \bar{b}_n \sigma_n}{b_m}\}} |\mathcal{D}_m|^2 \big| \mathcal{F}_{m-1}\big] = 0, \text{ in probability.}$$

   Then, it follows that

   $$\sigma_n^{-1}\Xi_n \overset{\text{stably}}{\Longrightarrow} H^{-1}\mathcal{N}(0,\Gamma), \text{ on } A,$$

   where the right-hand side stands for the random distribution being obtained when applying the $\mathcal{F}_\infty$-measurable matrix H^{-1} onto a normally distributed random variable with mean zero and covariance Γ.

The proof relies on two technical estimates taken from [7]. Based on a monotonically decreasing sequence $(\gamma_n)_{n\in\mathbb{N}}$ of strictly positive reals we define times $(t_n)_{n\in\mathbb{N}_0}$ via

$$t_n = \sum_{m=1}^{n} \gamma_m.$$

We cite [7, Lemma 2.3].

Lemma 5.2. *If $\lim_{n\to\infty} n\gamma_n = \infty$, then for every $C > 0$*

$$\lim_{n\to\infty} \frac{1}{n} \#\{l \in \{1,\dots,n\} : t_n - t_l \le C\} = 0.$$

We cite [7, Lemma 2.2].

Lemma 5.3. *We define for each $l \in \mathbb{N}$ the function $F_l : [0,\infty) \to [0,\infty)$ by demanding that for every $k \ge l$ and $s \in [t_{k-1} - t_{l-1}, t_k - t_{l-1})$*

$$F_l(s) = \frac{\gamma_l b_k}{\gamma_k b_l}.$$

If $\frac{b_{n+1}\gamma_n}{b_n\gamma_{n+1}} = 1 + o(\gamma_n)$, then

(i) F_l converges pointwise to 1

(ii) there exists a measurable function $\bar{F}$ and $n_0 \in \mathbb{N}$ such that $F_l \le \bar{F}$ for all $l \ge n_0$ and

$$\int_0^\infty \bar{F}(s)(s \vee 1)e^{-Ls} ds < \infty.$$

The following lemma is a slight variation of [7, Lemma 2.6].

Lemma 5.4. *Let $(H_n)_{n\in\mathbb{N}}$ be a (deterministic) sequence of symmetric matrices that converges to a matrix H with $\sigma(H) \subset (-\infty, 0)$. If $\frac{b_{n+1}\gamma_n}{b_n\gamma_{n+1}} = 1 + o(\gamma_n)$ then $\bar{\mathcal{H}}_n$ as defined in (5.1) satisfies*

$$\limsup_{l,n\to\infty,\, t_n - t_l \to\infty} \left\| \bar{\mathcal{H}}_n[l,n] + H^{-1} \right\| = 0.$$

Proof. Let $l, k \in \mathbb{N}_0$ with $l \le k$. We will first provide an estimate for $e^{(t_k - t_l)H_n} - \prod_{r=l+1}^{k}(\mathbb{1} + \gamma_r H_n)$ on the basis of the following telescoping sum representation:

$$e^{(t_k-t_l)H_n} - \prod_{r=l+1}^{k} (\mathbb{1} + \gamma_r H_n) = \sum_{q=l+1}^{k} e^{(t_{q-1}-t_l)H_n} (e^{\gamma_q H_n} - (\mathbb{1} + \gamma_q H_n)) \prod_{r=q+1}^{k} (\mathbb{1} + \gamma_r H_n). \tag{5.3}$$

Each term in the latter sum is a product of three matrices and we will analyse the norm of these individually.

We will use that the spectrum of a matrix depends continuously on the matrix. Let $\lambda^{(1)}, \dots, \lambda^{(d)}$ denote the eigenvalues of H. For $n \in \mathbb{N}$ one can enumerate the eigenvalues $\lambda_n^{(1)}, \dots, \lambda_n^{(d)}$ of H_n in such a way that $\lim_{n\to\infty} \lambda_n^{(i)} = \lambda^{(i)}$ for every $i = 1, \dots, d$ (see for instance [3, VI.1.4]). By assumption, $\sigma(H) \subset (-\infty, 0)$ so that there exist $C, L > 0, n_0 \in \mathbb{N}$ with

$$\sigma(H_n) \subset [-C, -L] \text{ for all } n \ge n_0.$$

Next note that for $\delta \ge 0$, $\mathbb{1} + \delta H_n$ has eigenvalues $1 + \delta\lambda_n^{(1)}, \dots, 1 + \delta\lambda_n^{(d)}$. These are all elements of the interval $[1 - \delta C, 1 - \delta L]$ and provided that $\delta \le 1/C$ we get that the spectral radius and likewise the matrix norm of $\mathbb{1} + \delta H_n$ are bounded by $1 - \delta L$. By

possibly increasing the value of n_0 we can guarantee that for all $k \geq n_0$, $\gamma_k \leq \frac{1}{C}$. For such n_0 we conclude that for all $k \geq l \geq n_0$ and $n \geq n_0$,

$$\|\mathcal{H}_n[l,k]\| = \Big\| \prod_{r=l+1}^{k} (\mathbb{1} + \gamma_r H_n) \Big\| \leq \prod_{r=l+1}^{k} \underbrace{(1-\gamma_r L)}_{\leq e^{-\gamma_r L}} \leq e^{-L(t_k - t_l)}. \tag{5.4}$$

Moreover, $e^{(t_k-t_l)H_n}$ has eigenvalues $\exp\{(t_k-t_l)\lambda_n^{(1)}\}, \dots, \exp\{(t_k-t_l)\lambda_n^{(d)}\}$ so that

$$\|e^{(t_k-t_l)H_n}\| \leq e^{-L(t_k-t_l)}.$$

Recall further that for a $d \times d$-matrix A

$$\|e^A - (\mathbb{1} + A)\| \leq \frac{1}{2} e^{\|A\|} \|A\|^2.$$

Altogether, we thus get with (5.3) that

$$\begin{aligned} \|e^{(t_k-t_l)H_n} - \mathcal{H}_n[l,k]\| &= \Big\| e^{(t_k-t_l)H_n} - \prod_{r=l+1}^{k} (\mathbb{1} + \gamma_r H_n) \Big\| \\ &\leq \frac{1}{2} e^{-(t_k-t_l)L + \gamma_1 L} e^{\gamma_1 \|H_n\|} \|H_n\|^2 \sum_{q=l+1}^{k} \gamma_q^2 \\ &\leq C' e^{-(t_k-t_l)L} \gamma_l (t_k - t_l) \end{aligned} \tag{5.5}$$

with $C' := \sup_{n \geq n_0} \|H_n\|^2 e^{\gamma_1(L + \|H_n\|)} \leq C^2 e^{\gamma_1(L+C)} < \infty$.

We note that, as H_n are symmetric matrices with $\sigma(H_n) \subset [-C, -L]$ for all $n \geq n_0$, H_n is invertible and

$$H_n^{-1} \to H^{-1}.$$

Therefore, it suffices to show that

$$\limsup_{l,n \to \infty,\ t_n - t_l \to \infty} \big\| \bar{\mathcal{H}}_n[l,n] + H_n^{-1} \big\| = 0.$$

To establish this we consider for $n \geq l \geq n_0$, $I_1 = I_1(l,n) = \bar{\mathcal{H}}_n[l,n]$,

$$I_2 = I_2(l,n) = \frac{\gamma_l}{b_l} \sum_{k=l}^{n} b_k e^{(t_k-t_l)H_n} \quad \text{and} \quad I_3 = I_3(l,n) = \sum_{k=l}^{n} \gamma_k e^{(t_k-t_l)H_n}$$

and omit the (l,n)-dependence in the notation.

We analyse $\|I_1 - I_2\|$. Using F_l as introduced in Lemma 5.3 we get with (5.5) that

$$\begin{aligned} \|I_1 - I_2\| &\leq \sum_{k=l}^{n} \frac{\gamma_l b_k}{b_l} \|\mathcal{H}_n[l,k] - e^{(t_k-t_l)H_n}\| \\ &\leq C' \gamma_l \sum_{k=l}^{n} \frac{\gamma_l b_k}{b_l \gamma_k} e^{-(t_k-t_l)L} (t_k - t_l) \gamma_k \\ &= C' \gamma_l \sum_{k=l}^{n} \int_{t_{k-1}-t_{l-1}}^{t_k - t_{l-1}} F_l(s) e^{-(t_k-t_l)L} (t_k - t_l)\, ds. \end{aligned}$$

Each integral is taken over an interval $(t_{k-1} - t_{l-1}, t_k - t_{l-1}]$ and for the respective s we get

$$t_k - t_l \leq t_{k-1} - t_{l-1} \leq s \quad \text{and} \quad t_k - t_l = t_k - t_{l-1} - \gamma_l \geq s - \gamma_l.$$

Thus,

$$\|I_1 - I_2\| \leq C' e^{\gamma_1 L} \gamma_l \int_0^{t_n - t_{l-1}} F_l(s) e^{-sL} s \, ds.$$

By Lemma 5.3, there exists an integrable majorant for the latter integrand. Hence, $\|I_1 - I_2\|$ is uniformly bounded and converges to zero as $l, n \to \infty$ with $l \leq n$.

We analyse $\|I_2 - I_3\|$. One has

$$I_2 - I_3 = \sum_{k=l}^{n} \Big(\frac{\gamma_l b_k}{b_l \gamma_k} - 1 \Big) \gamma_k e^{(t_k - t_l) H_n} = \sum_{k=l}^{n} \int_{t_{k-1} - t_{l-1}}^{t_k - t_{l-1}} \big(F_l(s) - 1 \big) e^{(t_k - t_l) H_n} \, ds$$

and using that $\|e^{(t_k - t_l) H_n}\| \leq e^{-L(t_k - t_l)}$ we argue as before to get that

$$\|I_2 - I_3\| \leq e^{\gamma_1 L} \int_0^{t_n - t_{l-1}} |F_l(s) - 1| \, e^{-Ls} \, ds.$$

Again there exists an integrable majorant. Hence, $\|I_2 - I_3\|$ is uniformly bounded and with dominated convergence and Lemma 5.3 we conclude that the latter integral converges to zero as $l, n \to \infty$ with $l \leq n$.

We analyse $\|I_3 + H_n^{-1}\|$. Using that $H_n^{-1} = -\int_0^\infty e^{sH_n} \, ds$ we write

$$I_3 + H_n^{-1} = \sum_{k=l}^{n} \int_{t_{k-1} - t_{l-1}}^{t_k - t_{l-1}} (e^{(t_k - t_l) H_n} - e^{sH_n}) \, ds - \int_{t_n - t_{l-1}}^{\infty} e^{sH_n} \, ds.$$

For $s \in (t_{k-1} - t_{l-1}, t_k - t_{l-1}]$

$$\begin{aligned} \|e^{(t_k - t_l) H_n} - e^{sH_n}\| &= \|e^{(t_k - t_l) H_n} (\mathbb{1} - e^{(s - (t_k - t_l)) H_n})\| \\ &\leq e^{-(t_k - t_l) L} (s - (t_k - t_l)) C e^{(s - (t_k - t_l)) C} \leq C'' e^{-Ls} \gamma_l \end{aligned}$$

with $C'' = C e^{(C+L)\gamma_1}$. Hence, we get with $\|e^{sH_n}\| \leq e^{-Ls}$

$$\|I_3 + H_n^{-1}\| \leq C'' \gamma_l \int_0^{t_n - t_{l-1}} e^{-Ls} \, ds + \int_{t_n - t_{l-1}}^{\infty} e^{-Ls} \, ds \leq \frac{C''}{L} \gamma_l + \frac{1}{L} e^{-L(t_n - t_{l-1})}.$$

Letting $l, n \to \infty$ with $t_n - t_l \to \infty$ the previous term tends to zero.

Altogether, it thus follows that

$$\limsup_{l, n \to \infty,\, t_n - t_l \to \infty} \underbrace{\big\| \bar{\mathcal{H}}_n[l, n] + H^{-1} \big\|}_{\leq \|I_1 - I_2\| + \|I_2 - I_3\| + \|I_3 + H_n^{-1}\| + \|H_n^{-1} - H^{-1}\|} = 0. \qquad \square$$

Lemma 5.5. *Let $L, C \in (0, \infty)$. If $\frac{b_{n+1}\gamma_n}{b_n \gamma_{n+1}} = 1 + o(\gamma_n)$ then there exist constants $\tilde{C} < \infty$ and $\tilde{N} \in \mathbb{N}$ such that for every symmetric matrix H with*

$$\sigma(H) \subset [-C, -L]$$

one has for every $l, n \in \mathbb{N}$ with $\tilde{N} \leq l \leq n$

$$\|\bar{\mathcal{H}}[l, n]\| \leq \tilde{C},$$

where we denote

$$\mathcal{H}[i, j] = \prod_{r=i+1}^{j} (\mathbb{1} + \gamma_r H) \;\text{ and }\; \bar{\mathcal{H}}[i, j] = \sum_{r=i}^{j} \frac{\gamma_i b_r}{b_i} \mathcal{H}[i, r].$$

Proof. By Lemma 5.3, there exists $\tilde{N} \in \mathbb{N}$ and a measurable function $\bar{F}$ such that $F_l \leq \bar{F}$ for all $l \geq \tilde{N}$ with

$$\int_0^\infty \bar{F}(s)e^{-Ls}ds < \infty.$$

By possibly increasing $\tilde{N}$ we can guarantee that $\gamma_l < \frac{1}{C}$ for all $l \geq \tilde{N}$. Note that estimate (5.4) prevails for arbitrary symmetric matrices H with $\sigma(H) \subset [-C, -L]$. Hence, for $l, n \in \mathbb{N}$ with $\tilde{N} \leq l \leq n$

$$\begin{aligned}\|\bar{\mathcal{H}}[l,n]\| &= \Big\|\sum_{k=l}^n \frac{\gamma_l b_k}{b_l}\mathcal{H}[l,k]\Big\| \leq \sum_{k=l}^n \frac{\gamma_l b_k}{b_l}e^{-L(t_k-t_l)\gamma_k} \\ &= \sum_{k=l}^n \int_{t_{k-1}-t_{l-1}}^{t_k-t_{l-1}} \frac{\gamma_l b_k}{b_l\gamma_k}e^{-L(t_k-t_l)\gamma_k}\,ds \leq e^{\gamma_1 L}\int_0^{t_n-t_{l-1}} F_l(s)e^{-sL}ds \\ &\leq e^{\gamma_1 L}\int_0^\infty \bar{F}(s)e^{-Ls}ds < \infty,\end{aligned}$$

which proves uniform boundedness. □

We are now in the position to prove the main result of this section.

Proof of Theorem 5.1. For $N \in \mathbb{N}$, $L, C \in (0,\infty)$ and $n \geq N$ we consider the events

$$\mathbb{A}_{N..n,C,L} = \{\sigma(H_m) \subset [-C,-L] \text{ for } m = N,\dots,n\} \text{ and } \mathbb{A}_{N..\infty,C,L} = \bigcap_{m\geq N} \mathbb{A}_{N..m,C,L}.$$

We will use Theorem A.6 to verify the statement on the event $\mathbb{A}_{N..\infty,C,L} \cap A$. By assumption, $H_n \to H$, almost surely, on A, so that in particular, almost surely, on A,

$$\min\sigma(H_n) \to \min\sigma(H) \text{ and } \max\sigma(H_n) \to \max\sigma(H) < 0.$$

Hence, up to nullsets,

$$A \subset \bigcup_{N,r,l\in\mathbb{N}} \mathbb{A}_{N..\infty,r,\frac{1}{l}}.$$

It thus suffices to prove the statement on $A \cap \mathbb{A}_{N..\infty,C,L}$ for fixed $N \in \mathbb{N}$ and $C, L > 0$, see Lemma A.3, and we briefly write for $n \geq N$

$$\mathbb{A}_n = \mathbb{A}_{N..n,C,L} \text{ and } \mathbb{A}_\infty = \mathbb{A}_{N..\infty,C,L}.$$

We denote by $\tilde{N}$ and $\tilde{C}$ the respective constants appearing in the statement of Lemma 5.5 and restrict attention to $n \in \mathbb{N}$ with $n_0(n) \geq \tilde{N} \vee N$. We will apply Theorem A.6 with $(Z_m^{(n)})_{m=1,\dots,n}$ given by

$$Z_m^{(n)} = \mathbb{1}_{\mathbb{A}_{n_0(n)}}\mathbb{1}_{\{m>n_0(n)\}}\frac{b_m}{\bar{b}_n\sigma_n}\bar{\mathcal{H}}_n[m,n]\,\mathcal{D}_m,$$

and with A and Γ in the Lemma replaced by $A\cap\mathbb{A}_\infty$ and $H^{-1}\Gamma(H^{-1})^\dagger$, respectively. Once we have verified that Theorem A.6 is applicable we conclude that, on $A \cap \mathbb{A}_\infty$,

$$\frac{1}{\sigma_n}\Xi_n = \sum_{m=1}^n Z_m^{(n)} \overset{\text{stably}}{\Longrightarrow} H^{-1}\mathcal{N}(0,\Gamma)$$

which finishes the proof.

It remains to verify the assumptions of Theorem A.6. For $m = 1,\dots,n_0(n)$ we have $Z_m^{(n)} = 0$ and, for $m = n_0(n)+1,\dots,n$, $\mathbb{1}_{\mathbb{A}_{n_0(n)}}\bar{\mathcal{H}}_n[m,n]$ is $\mathcal{F}_{n_0(n)}$-measurable and

hence $\mathcal{F}_{m-1}$-measurable, and the respective matrix norm is uniformly bounded by $\tilde{C}$. Consequently, $(Z_m^{(n)})_{m=1,\dots,n}$ is a sequence of martingale differences satisfying for $\varepsilon > 0$

$$\sigma_n^{-2} \sum_{m=n_0(n)+1}^{n} \mathbb{E}\Big[\mathbb{1}_{\mathbb{A}_{n_0(n)}} \mathbb{1}_{\{|\frac{b_m}{\bar{b}_n}\bar{\mathcal{H}}_n[m,n]\mathcal{D}_m|/\sigma_n \geq \varepsilon\}} \Big| \frac{b_m}{\bar{b}_n}\bar{\mathcal{H}}_n[m,n]\mathcal{D}_m\Big|^2 \Big| \mathcal{F}_{m-1}\Big]$$
$$\leq \tilde{C}^2\sigma_n^2 \sum_{m=n_0(n)+1}^{n} \frac{b_m^2}{\bar{b}_n^2}\mathbb{E}\Big[\mathbb{1}_{\{|\mathcal{D}_m| \geq \frac{\varepsilon \bar{b}_n \sigma_n}{\tilde{C} b_m}\}}|\mathcal{D}_m|^2 \Big| \mathcal{F}_{m-1}\Big]$$

and the latter term tends to zero in probability on A, by assumption.

It remains to control the asymptotics of

$$V_n := \sum_{m=1}^{n} \mathrm{cov}(Z_m^{(n)}|\mathcal{F}_{m-1})$$
$$= \sigma_n^{-2} \sum_{m=n_0(n)+1}^{n} \frac{b_m^2\delta_m^2}{\bar{b}_n^2}\mathbb{1}_{\mathbb{A}_{n_0(n)}}\bar{\mathcal{H}}_n[m,n]\mathrm{cov}(\delta_m^{-1}\mathcal{D}_m|\mathcal{F}_{m-1})\bar{\mathcal{H}}_n[m,n]^\dagger$$

on $A \cap \mathbb{A}_\infty$. By Lemma 5.2, we can choose a sequence $(L(n))$ such that $n_0(n) \leq L(n)$,

$$t_n - t_{L(n)} \to \infty \text{ and } n - L(n) = o(n).$$

Now, by assumption,

$$\sum_{m=L(n)+1}^{n} (b_m\delta_m)^2 = o((\sigma_n\bar{b}_n)^2).$$

As consequence of Assumption (3.a)

$$\kappa := \sup_{m \geq N} \big\|\mathrm{cov}(\delta_m^{-1}\mathcal{D}_m|\mathcal{F}_{m-1})\big\|$$

is almost surely finite on A. We thus get that on $A \cap \mathbb{A}_\infty$

$$\Big\| \sum_{m=L(n)+1}^{n} \big(\mathrm{cov}(Z_m^{(n)}|\mathcal{F}_{m-1}) - \sigma_n^{-2}\frac{b_m^2\delta_m^2}{\bar{b}_n^2}H^{-1}\Gamma(H^{-1})^\dagger\big)\Big\|$$
$$\leq \sigma_n^{-2} \sum_{m=L(n)+1}^{n} \frac{b_m^2\delta_m^2}{\bar{b}_n^2}\Big(\big\|\bar{\mathcal{H}}_n[m,n]\,\mathrm{cov}(\delta_m^{-1}\mathcal{D}_m|\mathcal{F}_{m-1})\,\bar{\mathcal{H}}_n[m,n]^\dagger\big\| + \big\|H^{-1}\Gamma(H^{-1})^\dagger\big\|\Big)$$
$$\leq 2\kappa\tilde{C}^2(\sigma_n\bar{b}_n)^{-2} \sum_{m=L(n)+1}^{n} (b_m\delta_m)^2 \to 0, \text{ almost surely, on } A \cap \mathbb{A}_\infty. \tag{5.6}$$

By assumption,

$$\rho_n := \sup_{m=n_0(n)+1,\dots,n} \big\|\mathrm{cov}(\delta_m^{-1}\mathcal{D}_m|\mathcal{F}_{m-1}) - \Gamma\big\| \to 0, \text{ almost surely, on } A,$$

and, by Lemma 5.4,

$$\rho_n' := \sup_{m=n_0(n)+1,\dots,L(n)} \big\|\bar{\mathcal{H}}_n[m,n] + H^{-1}\big\| \to 0, \text{ almost surely, on } A.$$

Consequently, one has for $m = n_0(n)+1, \dots, L(n)$, on $A \cap \mathbb{A}_\infty$,

$$\|\bar{\mathcal{H}}_n[m,n]\mathrm{cov}(\delta_i^{-1}\mathcal{D}_m|\mathcal{F}_{m-1})\bar{\mathcal{H}}_n[m,n]^\dagger - H^{-1}\Gamma(H^{-1})^\dagger\|$$
$$\leq \|\bar{\mathcal{H}}_n[m,n] + H^{-1}\|\|\mathrm{cov}(\delta^{-1}\mathcal{D}_m|\mathcal{F}_{m-1})\|\|\bar{\mathcal{H}}_n[m,n]^\dagger\|$$
$$+ \|H^{-1}\|\|\mathrm{cov}(\delta_m^{-1}\mathcal{D}_m|\mathcal{F}_{m-1}) - \Gamma\|\|\bar{\mathcal{H}}_n[m,n]^\dagger\| + \|H^{-1}\|\|\Gamma\|\|\bar{\mathcal{H}}_n[m,n]^\dagger + (H^{-1})^\dagger\|$$
$$\leq 2\kappa\tilde{C}\rho_n' + \tilde{C}^2\rho_n$$

and thus, on $A \cap \mathbb{A}_\infty$,

$$\Big\| \sum_{i=n_0(n)+1}^{L(n)} \big(\mathrm{cov}(Z_i^{(n)}|\mathcal{F}_{i-1}) - \sigma_n^{-2} \frac{b_i^2\delta_i^2}{\bar{b}_n^2} H^{-1}\Gamma(H^{-1})^\dagger\big)\Big\| \leq 2\kappa\tilde{C}\rho_n' + \tilde{C}^2\rho_n \to 0, \text{ a.s.}$$

By definition, one has $\sigma_n^{-2}\sum_{i=n_0(n)+1}^{n} \frac{b_i^2\delta_i^2}{\bar{b}_n^2} = 1$ so that together with (5.6) we obtain that on $A \cap \mathbb{A}_\infty$

$$\big\|V_n - H^{-1}\Gamma(H^{-1})^\dagger\big)\big\| \to 0, \text{ almost surely.}$$

This finishes the proof. □

Remark 5.6. Theorem 5.1 remains true when replacing $(\sigma_n)_{n\in\mathbb{N}}$ by $(\sigma_n')_{n\in\mathbb{N}}$ given by

$$\sigma_n' = \frac{1}{\bar{b}_n}\sqrt{\sum_{l=1}^{n}(b_l\delta_l)^2}$$

and $(n_0(n))_{n\in\mathbb{N}}$ being a sequence with

$$\sum_{i=1}^{n_0(n)} b_i^2\delta_i^2 = o\Big(\sum_{i=1}^{n} b_i^2\delta_i^2\Big).$$

Indeed, in that case we have

$$\frac{\sigma_n^2}{(\sigma_n')^2} = 1 - \frac{\sum_{i=1}^{n_0(n)} b_i^2\delta_i^2}{\sum_{i=1}^{n} b_i^2\delta_i^2} \to 1.$$

6 Technical preliminaries

In this section, we provide some technical estimates. First, we deduce that the notion of a regular function entails certain Taylor type error estimates. Technically, we need to take care of the fact that segments connecting two points are not necessarily contained in the domain of the function.

Lemma 6.1. *Let $U \subset \mathbb{R}^d$ be an open and bounded set, $g : U \to \mathbb{R}^d$ be a mapping and $\alpha_g \in (0, 1]$.*

(i) If g has regularity α_g, then for every $\delta > 0$ there exists a constant C_g such that for all $x, y \in U_\delta = \{z \in U : d(z, U^c) > \delta\}$

(a) $|g(x)| \vee \|Dg(x)\| \leq C_g$,
(b) $|g(y) - (g(x) + Dg(x)(y - x))| \leq C_g|y - x|^{1+\alpha_g}$ and
(c) $\|Dg(y) - Dg(x)\| \leq C_g|y - x|^{\alpha_g}$.

(ii) If $g : U \to \mathbb{R}^d$ has regularity α_g around a subset $M \subset \mathbb{R}^d$, then there exists a constant C_g such that for all $x \in M \cap U$ and $y \in U$

(a) $|g(x)| \vee \|Dg(x)\| \leq C_g$ and
(b) $|g(y) - (g(x) + Dg(x)(y - x))| \leq C_g|y - x|^{1+\alpha_g}$

and for all $x, y \in M \cap U$

(c) $\|Dg(y) - Dg(x)\| \leq C_g|y - x|^{\alpha_g}$.

Proof. **(i):** g is continuous and thus bounded on the compact set $\overline{U_\delta} \subset U$ so that properties (a) and (c) follow from the Hölder continuity of Dg and the boundedness of U. By Taylor's formula property (b) holds for every $x, y \in U$ with the constant $\sup \|Dg\|$ whenever the segment connecting x and y lies in U. Now suppose that properties (a) and (c) are true for the constant C and that $\sup_{x,y\in U} d(x,y) \leq C$. We consider two points $x, y \in U_\delta$ whose segment is *not* contained in U. Then we have that $d(x,y) \geq 2\delta$ so that

$$\begin{aligned}|g(y) - (g(x) + Dg(x)(y-x))| &\leq |g(y)| + |g(x)| + |Dg(x)(y-x))| \\ &\leq 2C + C^2 \leq \frac{2C + C^2}{(2\delta)^{1+\alpha_g}} |y-x|^{1+\alpha_g}.\end{aligned}$$

Consequently, properties (a), (b) and (c) are true on U_δ for a sufficiently large constant C_g.

(ii): Note that properties (b) and (c) are true for a sufficiently large constant and that (a) follows with the boundedness of U and (b). □

Let now U denote an (F, M)-attractor with stability $L > 0$ and bound C and suppose that $\Phi : U_\Phi \to U$ is a nice representation for M on U of regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$ with $\alpha_f, \alpha_\Phi, \alpha_\Psi \in (0, 1]$. We fix $\delta > 0$ and choose $\rho \in (0, \delta/4]$ as in (i) of Lemma 3.6 and again denote by U_δ^ρ the set

$$U_\delta^\rho = \bigcup_{\substack{y\in M: \\ d(y,U^c)>\delta}} B_\rho(y).$$

Recall that by Lemma 3.6, for every $x \in U_\delta^\rho$ there exists a unique closest element x^* in M and one has

$$x^* = \Phi(\Psi_\zeta(x), 0) \in U \cap M.$$

Now let (X_n) and (γ_n) as introduced in (1.1). We analyse the dynamical system based on the nice representation introduced above. That means, for every $n \in \mathbb{N}$, we define on the event $\{X_n \in U\}$ the coordinates

$$\begin{pmatrix}\zeta_n \\ \theta_n\end{pmatrix} = \Psi(X_n) = \begin{pmatrix}\Psi_\zeta(X_n) \\ \Psi_\theta(X_n)\end{pmatrix}, \tag{6.1}$$

where

$$\Psi_\zeta(x) = \begin{pmatrix}\Psi_1(x) \\ \vdots \\ \Psi_{d_\zeta}(x)\end{pmatrix} \quad \text{and} \quad \Psi_\theta(x) = \begin{pmatrix}\Psi_{d_{\zeta+1}}(x) \\ \vdots \\ \Psi_d(x)\end{pmatrix}.$$

Crucial in our approach is the analysis of a linearised system. For a fixed element $\bar{x} = \Phi(\bar{\zeta}, 0) \in M \cap U$ and every $n \in \mathbb{N}$ we define on the event that X_{n-1} and X_n both are in U the random variable Υ_n via

$$\begin{pmatrix}\zeta_n \\ \theta_n\end{pmatrix} = \begin{pmatrix}\zeta_{n-1} \\ \theta_{n-1}\end{pmatrix} + \gamma_n\Big(\begin{pmatrix}0 \\ H_{\bar{x}}\theta_{n-1}\end{pmatrix} + \Upsilon_n\Big), \tag{6.2}$$

where $H_{\bar{x}}$ is the matrix with

$$H_{\bar{x}}\theta = D\Psi_\theta(\bar{x}) Df(\bar{x}) (D\Psi(\bar{x}))^{-1} \begin{pmatrix}0 \\ \theta\end{pmatrix}.$$

Informally,

$$\Upsilon_n = D\Psi(X_{n-1}) D_n + \text{error term}$$

and we control the error term in the following lemma.

Lemma 6.2. *Suppose that $\Phi : U_\Phi \to U$ is a nice representation for M on a bounded and open set U with regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi) \in (0,1]^3$. Let $\delta > 0$ and $\rho \in (0, \delta/4]$ as in (i) of Lemma 3.6. There exists a constant $\tilde{C} \geq 0$ such that the following is true. If for $x \in U_\delta^\rho$, $\gamma \in (0, \gamma_0]$, $u \in \mathbb{R}^d$ one has*

$$x' := x + \gamma(f(x) + u) \in U_\delta,$$

then for every $\bar{x} \in M \cap U$, $\theta = \Psi_\theta(x)$ and $\Upsilon \in \mathbb{R}^d$ given by

$$\Psi(x') - \Psi(x) = \gamma\Big(\begin{pmatrix} 0 \\ H_{\bar{x}}\theta \end{pmatrix} + \Upsilon\Big), \text{ where } H_{\bar{x}}\theta = D\Psi_\theta(\bar{x})Df(\bar{x})(D\Psi(\bar{x}))^{-1}\begin{pmatrix} 0 \\ \theta \end{pmatrix}, \quad (6.3)$$

one has

$$|\Upsilon - D\Psi(x)u| \leq \tilde{C}\big(\gamma^{\alpha_\Psi}|u|^{1+\alpha_\Psi} + d(x,M)d(x,\bar{x})^\alpha\big),$$

where $\alpha = \alpha_\Psi \wedge \alpha_f \wedge \alpha_\Phi$.

Proof. Note that, by assumption, x, x' and x^* are all in $U_{\delta/2}$ and we will use the Taylor-type estimates of Lemma 6.1 without further mentioning. For $\bar{x} \in M \cap U$ we set $\bar{H}_{\bar{x}} = D\Psi(\bar{x})Df(\bar{x})(D\Psi(\bar{x}))^{-1}$. Then,

$$\begin{pmatrix} 0 \\ H_{\bar{x}}\theta \end{pmatrix} = \bar{H}_{\bar{x}}\begin{pmatrix} 0 \\ \theta \end{pmatrix}$$

since a vector $(0,\theta)^\dagger$ is mapped by $(D\Psi(\bar{x}))^{-1} = D\Phi(\Psi(\bar{x}))$ to a vector in $N_z M$ which is mapped itself by $Df(z)$ to a vector in $N_z M$ (see Remark 2.2) and then by $D\Psi(z)$ to a vector in $\{0\}^{d_\zeta} \times \mathbb{R}^{d_\theta}$. As consequence of (6.3) we get that

$$\Upsilon = \frac{1}{\gamma}(\Psi(x + \gamma(f(x) + u)) - \Psi(x)) - \bar{H}_{\bar{x}}\begin{pmatrix} 0 \\ \theta \end{pmatrix}.$$

Using the α_Ψ-regularity of Ψ we get that

$$\frac{1}{\gamma}(\Psi(x + \gamma(f(x) + u)) - \Psi(x)) = D\Psi(x)(f(x) + u) + \mathcal{O}(\gamma^{\alpha_\Psi}(|f(x)|^{1+\alpha_\Psi} + |u|^{1+\alpha_\Psi})).$$

Here and elsewhere in the proof all $\mathcal{O}$-terms are uniform over all allowed choices of x, x', $\bar{x}$ and γ. By Lemma 3.6, x has a unique closest M-element $x^* \in U \cap M$ and using the α_f regularity of f and the boundedness of U we get that

$$|f(x)| = |f(x) - f(x^*)| = \mathcal{O}(|x - x^*|) = \mathcal{O}(d(x,M)).$$

Hence,

$$\frac{1}{\gamma}(\Psi(x + \gamma(f(x) + u)) - \Psi(x)) = D\Psi(x)(f(x) + u) + \mathcal{O}(\gamma^{\alpha_\Psi}(d(x,M)^{1+\alpha_\Psi} + |u|^{1+\alpha_\Psi})). \quad (6.4)$$

and with the α_Ψ-regularity of Ψ we get

$$D\Psi(x)f(x) = D\Psi(x^*)f(x) + \mathcal{O}(|x - x^*|^{\alpha_\Psi}|f(x)|) = D\Psi(x^*)f(x) + \mathcal{O}(d(x,M)^{1+\alpha_\Psi}). \quad (6.5)$$

Furthermore, Lemma 3.6 yields that $|\theta| = d(x,M)$, so that with $f(x^*) = 0$

$$\begin{aligned} D\Psi(x^*)f(x) &= D\Psi(x^*)Df(x^*)(x - x^*) + \mathcal{O}(d(x,M)^{1+\alpha_f}) \\ &= \underbrace{D\Psi(x^*)Df(x^*)(D\Psi(x^*))^{-1}}_{=\bar{H}_{x^*}}\begin{pmatrix} 0 \\ \theta \end{pmatrix} + \mathcal{O}(d(x,M)^{1+\alpha_f} + d(x,M)^{1+\alpha_\Phi}). \end{aligned} \quad (6.6)$$

Insertion of (6.4), (6.5) and (6.6) into the above representation of Υ gives together with the uniform boundedness of γ and $d(x,M)$

$$\Upsilon = D\Psi(x)u + (\bar{H}_{x^*} - \bar{H}_{\bar{x}}) \begin{pmatrix} 0 \\ \theta \end{pmatrix} + \mathcal{O}(d(x,M)^{1+\alpha} + \gamma^{\alpha_\Psi}|u|^{1+\alpha_\Psi}).$$

On the relevant domains $D\Psi$, Df and $D\Phi$ are Hölder continuous with parameter α and uniformly bounded so that $\|\bar{H}_{x^*} - \bar{H}_{\bar{x}}\| = \mathcal{O}(|x^* - \bar{x}|^\alpha)$. Since $|\theta| = d(x,M)$ we finally get that

$$\begin{aligned}\Upsilon &= D\Psi(x)u + \mathcal{O}(d(x,M)|\bar{x} - x^*|^\alpha + d(x,M)^{1+\alpha} + \gamma^{\alpha_\Psi}|u|^{1+\alpha_\Psi}) \\ &= D\Psi(x)u + \mathcal{O}(d(x,M)|x - \bar{x}|^\alpha + \gamma^{\alpha_\Psi}|u|^{1+\alpha_\Psi}).\end{aligned}$$ □

Proposition 6.3. *Let U be an (F,M)-attractor with stability $L > 0$ and bound C and suppose that $\Phi : U_\Phi \to U$ is a nice representation for M on U with regularity $(\alpha_f, \alpha_\Phi, \alpha_\Psi) \in (0,1]^3$. Set $\alpha = \alpha_\Psi \wedge \alpha_f \wedge \alpha_\Phi$. Let $(X_n)_{n\in\mathbb{N}_0}$ be as in (1.1) satisfying the following assumptions:*

- *$(\mathbb{1}_U(X_{n-1})D_n)_{n\in\mathbb{N}}$ is a sequence of square-integrable martingale differences,*
- *$(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals with $\gamma_n \to 0$ and $\sum \gamma_n = \infty$,*
- *$(\sigma_n^{\mathrm{RM}})_{n\in\mathbb{N}}$ is a sequence of strictly positive reals with*

$$L'' := \limsup_{n\to\infty} \frac{1}{\gamma_n} \frac{\sigma_{n-1}^{\mathrm{RM}} - \sigma_n^{\mathrm{RM}}}{\sigma_n^{\mathrm{RM}}} < L$$

and

$$\limsup_{n\to\infty} \Big(\frac{\sigma_n^{\mathrm{RM}}}{\sqrt{\gamma_n}}\Big)^{-1} \mathbb{E}\big[\mathbb{1}_U(X_{n-1})|D_n|^2\big]^{1/2} < \infty. \tag{6.7}$$

Let $\delta > 0$ and $\rho \in (0, \delta/4]$ be as in (i) of Lemma 3.6 and suppose that inequality (3.3) of Prop. 3.5 is true on U_δ^ρ for a $L' \in (L'', L)$ that is $d(x + \gamma f(x), M) \le (1 - \gamma L')d(x,M)$ for all $x \in U_\delta^\rho$ and $\gamma \in [0, C^{-1}]$. Then for every $N \in \mathbb{N}$, as $n \to \infty$,

$$\sup_{m=n_0(n)+1,\dots,n} |\zeta_m - \zeta_{n_0(n)}| = \mathcal{O}_P\big(\varepsilon_n^{\mathrm{RM}}\big) \text{ , on } \mathbb{U}_{N..\infty}^{\delta,\rho},$$

where ζ_m ($m \in \mathbb{N}_0$) is well-defined via (6.1) on $\{X_m \in U\}$,

$$\varepsilon_n^{\mathrm{RM}} = \sum_{k=n_0(n)+1}^{n} \big((\sqrt{\gamma_k}\sigma_k^{\mathrm{RM}})^{1+\alpha_\Psi} + \gamma_k(\sigma_{k-1}^{\mathrm{RM}})^{1+\alpha}\big) + \sqrt{\sum_{k=n_0(n)+1}^{n} \gamma_k(\sigma_k^{\mathrm{RM}})^2} \tag{6.8}$$

and

$$\mathbb{U}_{N..n}^{\delta,\rho} = \{\forall l = N,\dots,n : X_l \in U_\delta^\rho\} \text{ and } \mathbb{U}_{N..\infty}^{\delta,\rho} = \bigcap_{n\ge N} \mathbb{U}_{N..n}^{\delta,\rho}.$$

Proof. By Theorem 4.1, there exists a constant $\tilde{C} \in (0,\infty)$ such that for all $N \in \mathbb{N}$

$$\limsup_{n\to\infty} (\sigma_n^{\mathrm{RM}})^{-1} \mathbb{E}\big[\mathbb{1}_{\mathbb{U}_{N..n-1}^{\delta,\rho}} d(X_n, M)^2\big]^{1/2} \le \tilde{C}. \tag{6.9}$$

We fix $N \in \mathbb{N}$ and briefly write $\mathbb{U}_k = \mathbb{U}_{N..k}^{\delta,\rho}$ for $k \ge N$. By choice of ρ, Lemma 6.2 is applicable on U_δ^ρ and we conclude that for all m for which X_{m-1} and X_m lie in U_δ^ρ we have

$$\zeta_m = \zeta_{m-1} + \gamma_m D\Psi_\zeta(X_{m-1})D_m + \mathcal{O}(\gamma_m^{1+\alpha_\Psi}|D_m|^{1+\alpha_\Psi} + \gamma_m d(X_{m-1}, M)^{1+\alpha}).$$

Here we used the lemma with $x = X_{m-1}$, $x' = X_m$, $\bar{x} = X^*_{m-1}$ and $\gamma = \gamma_m$. Note that the $\mathcal{O}$-term is uniformly bounded over all realisations and allowed choices of m.

We consider $n \in \mathbb{N}$ with $n_0(n) \geq N$. On $\mathbb{U}_n$, one has for $m = n_0(n)+1, \dots, n$,

$$\begin{aligned}&\zeta_m - \zeta_{n_0(n)}\\ &= \underbrace{\sum_{k=n_0(n)+1}^{m} \gamma_k D\Psi_\zeta(X_{k-1})D_k}_{=:A^{(1)}_m} + \underbrace{\mathcal{O}\Big(\sum_{k=n_0(n)+1}^{n} \gamma_k^{1+\alpha_\Psi}|D_k|^{1+\alpha_\Psi} + \gamma_k d(X_{k-1},M)^{1+\alpha}\Big)}_{=:A^{(2)}_m}.\end{aligned}$$

For ease of notation we omit the n-dependence in the notation of the A-terms. We control

$$S^{(i)}_n := \sup_{m=n_0(n)+1,\dots,n} \big|A^{(i)}_m\big|$$

for the two choices of i separately.

By the boundedness of $D\Psi_\zeta$, the sequence $(\mathbb{1}_{\mathbb{U}_{k-1}}\gamma_k D\Psi_\zeta(X_{k-1})D_k)_{k=n_0(n)+1,\dots,n}$ defines a sequence of square integrable martingale differences. Thus, we get with Doob's martingale inequality, the uniform boundedness of $D\Psi_\zeta$ and (6.7) that

$$\mathbb{E}[|\mathbb{1}_{\mathbb{U}_n} S^{(1)}_n|^2] \leq 4C_\Psi\, \mathbb{E}\Big[\sum_{k=n_0(n)+1}^{n} \mathbb{1}_{\mathbb{U}_{k-1}}\gamma_k^2|D_k|^2\Big] = \mathcal{O}\Big(\sum_{k=n_0(n)+1}^{n} \gamma_k(\sigma^{\mathrm{RM}}_k)^2\Big).$$

Hence,

$$S^{(1)}_n = \mathcal{O}_P\Big(\sqrt{\sum_{k=n_0(n)+1}^{n} \gamma_k(\sigma^{\mathrm{RM}}_k)^2}\Big), \text{ on } \mathbb{U}_\infty,$$

see Remark B.2. It remains to bound the second term.

Note that by assumption

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=n_0(n)+1}^{n} \mathbb{1}_{\mathbb{U}_{k-1}}\gamma_k^{1+\alpha_\Psi}|D_k|^{1+\alpha_\Psi}\Big] &\leq \sum_{k=n_0(n)+1}^{n} \gamma_k^{1+\alpha_\Psi}\mathbb{E}[\mathbb{1}_{\mathbb{U}_{k-1}}|D_k|^2]^{(1+\alpha_\Psi)/2}\\ &= \mathcal{O}\Big(\sum_{k=n_0(n)+1}^{n} \big(\sqrt{\gamma_k}\sigma^{\mathrm{RM}}_k\big)^{1+\alpha_\Psi}\Big)\end{aligned}$$

and, with (6.9),

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=n_0(n)+1}^{n} \gamma_k\mathbb{1}_{\mathbb{U}_{k-1}}d(X_{k-1},M)^{1+\alpha}\Big] &\leq \sum_{k=n_0(n)+1}^{n} \gamma_k\mathbb{E}\big[\mathbb{1}_{\mathbb{U}_{k-1}}d(X_{k-1},M)^2\big]^{(1+\alpha)/2}\\ &= \mathcal{O}\Big(\sum_{k=n_0(n)+1}^{n} \gamma_k(\sigma^{\mathrm{RM}}_{k-1})^{1+\alpha}\Big),\end{aligned}$$

so that (see again Remark B.2)

$$S^{(2)}_n = \mathcal{O}_P\Big(\sum_{k=n_0(n)+1}^{n} \big(\big(\sqrt{\gamma_k}\sigma^{\mathrm{RM}}_k\big)^{1+\alpha_\Psi} + \gamma_k(\sigma^{\mathrm{RM}}_{k-1})^{1+\alpha}\big)\Big).$$

Together with the respective bound for $S^{(1)}_N$ above, this finishes the proof of the proposition. □

Proposition 6.4. *We assume the same assumptions as in Proposition 6.3. Then, for every $N \in \mathbb{N}$, we have*

$$\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k(\gamma_k^{\alpha_\Psi}|D_k|^{1+\alpha_\Psi} + d(X_{k-1}, M)d(X_{k-1}, X^*_{n_0(n)})^\alpha) = \mathcal{O}_P\big(\varepsilon_n^{\mathrm{RP}}\big), \text{ on } \mathbb{U}^{\delta,\rho}_{N..\infty}, \tag{6.10}$$

where

$$\varepsilon_n^{\mathrm{RP}} = \frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k\big(\gamma_k^{-\frac{1-\alpha_\Psi}{2}}(\sigma_k^{\mathrm{RM}})^{1+\alpha_\Psi} + (\sigma_{k-1}^{\mathrm{RM}})^{1+\alpha} + \sigma_{k-1}^{\mathrm{RM}}(\varepsilon_n^{\mathrm{RM}})^\alpha\big).$$

Proof. Fix $N \in \mathbb{N}$, consider $n \in \mathbb{N}$ with $n_0(n) \geq N$ and briefly write $\mathbb{U}_k = \mathbb{U}^{\delta,\rho}_{N..k}$ for $k \geq N$. First note that with Lemma 3.6 and the convexity of U_Φ for $k > n_0(n)$, on $\mathbb{U}_\infty$,

$$\begin{aligned} |X_{k-1} - X^*_{n_0(n)}| &\leq |X_{k-1} - X^*_{k-1}| + |X^*_{k-1} - X^*_{n_0(n)}| \\ &\leq d(X_{k-1}, M) + C_\Phi|\zeta_{k-1}, \zeta_{n_0(n)})|. \end{aligned}$$

Using this inequality the left-hand side of (6.10) is transformed into the sum of three terms that we will analyse independently below.

1) *Analysis of the first term.* First, we provide an asymptotic bound for

$$\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k d(X_{k-1}, M)d(\zeta_{k-1}, \zeta_{n_0(n)})^\alpha,$$

on $\mathbb{U}_\infty$. By choice of ρ, we have validity of (4.5) and we get that

$$\begin{aligned} \mathbb{E}\Big[\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k \mathbb{1}_{\mathbb{U}_{k-1}} d(X_{k-1}, M)\Big] &\leq \frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k \mathbb{E}[\mathbb{1}_{\mathbb{U}_{k-1}} d(X_{k-1}, M)^2]^{1/2} \\ &= \mathcal{O}\Big(\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k \sigma_{k-1}^{\mathrm{RM}}\Big). \end{aligned}$$

Hence,

$$\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k \mathbb{1}_{\mathbb{U}_{k-1}} d(X_{k-1}, M) = \mathcal{O}_P\Big(\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k \sigma_{k-1}^{\mathrm{RM}}\Big), \text{ on } \mathbb{U}_\infty.$$

With Proposition 6.3 we conclude that, on $\mathbb{U}_\infty$,

$$\begin{aligned} &\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k d(X_{k-1}, M)d(\zeta_{k-1}, \zeta_{n_0(n)})^\alpha \\ &\quad \leq \Big(\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k d(X_{k-1}, M)\Big) \sup_{m=n_0(n)+1,\dots,n} |\zeta_m - \zeta_{n_0(n)}|^\alpha \\ &\quad = \mathcal{O}_P\Big(\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k \sigma_{k-1}^{\mathrm{RM}}(\varepsilon_n^{\mathrm{RM}})^\alpha\Big). \end{aligned}$$

2) *Analysis of the second term.* Second, we analyse

$$\frac{1}{\overline{b}_n} \sum_{k=n_0(n)+1}^{n} b_k d(X_{k-1}, M)^{1+\alpha}.$$

With (4.5) we get that

$$\mathbb{E}\Big[\frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k \mathbb{1}_{\mathbb{U}_{k-1}} d(X_{k-1},M)^{1+\alpha}\Big] \le \frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k \mathbb{E}[\mathbb{1}_{\mathbb{U}_{k-1}} d(X_{k-1},M)^2]^{(1+\alpha)/2}$$
$$= \mathcal{O}\Big(\frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k (\sigma_{k-1}^{\mathrm{RM}})^{1+\alpha}\Big),$$

so that

$$\frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k d(X_{k-1},M)^{1+\alpha} = \mathcal{O}_P\Big(\frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k (\sigma_{k-1}^{\mathrm{RM}})^{1+\alpha}\Big), \text{ on } \mathbb{U}_\infty.$$

3) *Analysis of the third term.* Similarly to before, we conclude that

$$\mathbb{E}\Big[\frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k \gamma_k^{\alpha_\Psi} \mathbb{1}_{\mathbb{U}_{k-1}} |D_k|^{1+\alpha_\Psi}\Big] \le \frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k \gamma_k^{\alpha_\Psi} \mathbb{E}[\mathbb{1}_{\{X_{k-1}\in U\}} |D_k|^2]^{(1+\alpha_\Psi)/2}$$
$$= \mathcal{O}\Big(\frac{1}{\bar{b}_n}\sum_{k=n_0(n)+1}^{n} b_k \gamma_k^{-\frac{1-\alpha_\Psi}{2}} (\sigma_k^{\mathrm{RM}})^{1+\alpha_\Psi}\Big)$$

with the obvious $\mathcal{O}_P$-bound on $\mathbb{U}_\infty$. The statement is obtained by combining the three estimates. □

7 The proofs of the main results

7.1 Proof of Theorem 2.9

Proof of Theorem 2.9. 1) *Feasible triples.* Let (U,δ,ρ) be a feasible triple in the sense of Proposition 3.7. We denote by $\mathbb{U}^{\mathrm{conv}} = \mathbb{U}^{\mathrm{conv}}_{\delta,\rho}$ the event that $(X_n)_{n\in\mathbb{N}_0}$ converges to some value in $M \cap U^\rho_\delta$. As explained in Remark 3.8, the statement of Theorem 2.9 follows once we showed stable convergence on $\mathbb{U}^{\mathrm{conv}}$.

2) *Separating the directions.* Recall that, by Lemma 3.6, for all $x \in U^\rho_\delta$ there is a unique closest M-element $x^* = \Phi(\Psi_\zeta(x),0) \in M \cap U^\rho_\delta$. For $m \in \mathbb{N}$ we define on the event $\{X_m \in U^\rho_\delta\}$ a random symmetric $d_\theta \times d_\theta$-matrix H_m via

$$H_m\theta = D\Psi_\theta(X_m^*) Df(X_m^*)(D\Psi(X_m^*))^{-1}\begin{pmatrix}0\\ \theta\end{pmatrix},$$

with symmetry following from Remark 2.2. For technical reasons, we set $H_m = 0$ on $\{X_m \in U^\rho_\delta\}^c$. Let $N \in \mathbb{N}$ and consider, for $m \ge N$, the events

$$\mathbb{U}_{N..m} := \{\forall l = N,\dots,m : X_l \in U^\rho_\delta\},$$
$$\mathbb{U}_{N..\infty} := \bigcap_{m'\ge N} \mathbb{U}_{N..m'} \text{ and } \mathbb{U}^{\mathrm{conv}}_{N..\infty} := \mathbb{U}^{\mathrm{conv}} \cap \mathbb{U}_{N..\infty}.$$

Note that

$$\mathbb{U}^{\mathrm{conv}} = \bigcup_{N\in\mathbb{N}} \mathbb{U}^{\mathrm{conv}}_{N..\infty}$$

so that as consequence of Lemma A.3 it suffices to prove stable convergence on $\mathbb{U}^{\mathrm{conv}}_{N..\infty}$ for arbitrarily fixed $N \in \mathbb{N}$. Note that, on $\mathbb{U}^{\mathrm{conv}}$, (H_l) converges to the symmetric random matrix H with

$$H\theta = D\Psi_\theta(X_\infty) Df(X_\infty)(D\Psi(X_\infty))^{-1}\begin{pmatrix}0\\ \theta\end{pmatrix}.$$

Set $A = D\Psi_\theta(X_\infty)\big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1}\Pi_{N_{X_\infty}M}$ and note that by monotonicity it suffices to consider large N. We briefly write

$$\mathbb{U}_m = \mathbb{U}_{N..m},\ \mathbb{U}_\infty = \mathbb{U}_{N..\infty} \text{ and } \mathbb{U}_\infty^{\mathrm{conv}} = \mathbb{U}_{N..\infty}^{\mathrm{conv}}.$$

In the following, we restrict attention to $n \in \mathbb{N}$ with $n_0(n) \geq N$ and consider $m \geq n_0(n)$. Note that

$$\bar\theta_m = \frac{1}{\bar b_m}\sum_{k=n_0(n)}^{m} b_k\theta_k \text{ and } \bar\zeta_m = \frac{1}{\bar b_m}\sum_{k=n_0(n)}^{m} b_k\zeta_k$$

are well-defined on $\mathbb{U}_m$, where (θ_k) and (ζ_k) are given by (6.1). Moreover, for $m > n_0(n)$, we set on $\mathbb{U}_m$

$$\Upsilon_m^{(n)} = \frac{1}{\gamma_m}(\Psi(X_m) - \Psi(X_{m-1})) - \begin{pmatrix} 0 \\ H_{n_0(n)}\theta_{m-1}\end{pmatrix}$$

and on $\mathbb{U}_m^c$, $\Upsilon_m^{(n)} = 0$. Now, on $\mathbb{U}_m$,

$$\theta_m = \theta_{m-1} + \gamma_m(H_{n_0(n)}\theta_{m-1} + \pi_\theta(\Upsilon_m^{(n)})), \tag{7.1}$$

so that by the variation of constant formula

$$\theta_m = \mathcal{H}_{n_0(n)}[n_0(n), m]\theta_{n_0(n)} + \sum_{\ell=n_0(n)+1}^{m}\gamma_l\mathcal{H}_{n_0(n)}[\ell, m]\pi_\theta(\Upsilon_l^{(n)}),$$

with $\mathcal{H}_{n_0(n)}[i,j]$ and $\bar{\mathcal{H}}_{n_0(n)}[i,j]$ ($i,j \in \mathbb{N}$ with $i \leq j$) being defined as in (5.1). Consequently, on $\mathbb{U}_n$,

$$\bar\theta_n = \frac{b_{n_0(n)}}{\bar b_n\gamma_{n_0(n)}}\bar{\mathcal{H}}_{n_0(n)}[n_0(n), n]\theta_{n_0(n)} + \frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\,\bar{\mathcal{H}}_{n_0(n)}[m,n]\,\pi_\theta(\Upsilon_m^{(n)}) \tag{7.2}$$

with the right-hand side being a random variable that is defined on the whole space Ω and we take the previous formula as definition of the random variable $\bar\theta_n$ outside of $\mathbb{U}_n$.

3) *Approximation by the linear system of Section 5.* We set

$$\Xi_n := \frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\,\bar{\mathcal{H}}_{n_0(n)}[m,n]\,\mathcal{D}_m$$

with $\mathcal{D}_m = \mathbb{1}_{\mathbb{U}_{m-1}}D\Psi_\theta(X_{m-1})D_m$. By Lemma 6.2, there exists a constant $\tilde C_1$ such that, on $\mathbb{U}_n$, for all $n_0(n) \leq m \leq n$

$$|\pi_\theta\Upsilon_m^{(n)} - \mathcal{D}_m| \leq \tilde C_1\big(\gamma_m^{\alpha_\Psi}|D_m|^{1+\alpha_\Psi} + d(X_{m-1}, M)d(X_{m-1}, X^*_{n_0(n)})^\alpha\big).$$

Assuming that N is sufficiently large, Lemma 5.5 yields existence of a constant $\tilde C_2$ such that, on $\mathbb{U}_n$,

$$\begin{aligned}&\Big|\Xi_n - \frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\,\bar{\mathcal{H}}_{n_0(n)}[m,n]\,\pi_\theta(\Upsilon_m^{(n)})\Big| \\ &\leq \tilde C_1\tilde C_2\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\big(\gamma_m^{\alpha_\Psi}|D_m|^{1+\alpha_\Psi} + d(X_{m-1}, M)d(X_{m-1}, X^*_{n_0(n)})^\alpha\big).\end{aligned}$$

By Proposition 6.4, the latter term is of order $\mathcal{O}_P\big(\varepsilon_n^{\mathrm{RP}}\big)$ on $\mathbb{U}_\infty$. Thus, assumption (2.14) guarantees that the previous error term is of order $o_P(\sigma_n)$ on $\mathbb{U}_\infty$.

4) *Analysis of* Ξ_n. Recall that on $\mathbb{U}_\infty^{\mathrm{conv}}$, one has $\lim_{n\to\infty} H_{n_0(n)} \to H$ with H satisfying

$$H\theta = D\Psi_\theta(X_\infty)Df(X_\infty)(D\Psi(X_\infty))^{-1}\begin{pmatrix}0\\ \theta\end{pmatrix}.$$

By assumption, $Df(X_\infty)$ as a linear mapping from $N_{X_\infty}M$ to $N_{X_\infty}M$ is invertible and we get with elementary linear algebra that for $\theta \in \mathbb{R}^{d_\theta}$

$$H^{-1}\theta = D\Psi_\theta(X_\infty)\big(Df(X_\infty)|_{N_{X_\infty}M}\big)^{-1}(D\Psi(X_\infty))^{-1}\begin{pmatrix}0\\ \theta\end{pmatrix}.$$

Note that $(\mathcal{D}_m)_{m\geq N+1}$ is a sequence of martingale differences and one has, on $\mathbb{U}_\infty^{\mathrm{conv}}$, for $m > N$,

$$\begin{aligned}\mathrm{cov}((\delta_m^{\mathrm{diff}})^{-1}\mathcal{D}_m|\mathcal{F}_{m-1}) &= D\Psi_\theta(X_{m-1})(\delta_m^{\mathrm{diff}})^{-2}\mathrm{cov}(D_m|\mathcal{F}_{m-1})D\Psi_\theta(X_{m-1})^\dagger\\ &\to D\Psi_\theta(X_\infty)\,\Gamma\,D\Psi_\theta(X_\infty)^\dagger, \text{ almost surely.}\end{aligned}$$

Moreover, assumption (2.10) implies that for every $\varepsilon > 0$, on $\mathbb{U}_\infty^{\mathrm{conv}}$,

$$\begin{aligned}&\sigma_n^{-2}\sum_{m=n_0(n)+1}^{n}\frac{b_m^2}{\bar b_n^2}\mathbb{E}\big[\mathbb{1}_{\{|\mathcal{D}_m|>\frac{\varepsilon\bar b_n\sigma_n}{b_m}\}}|\mathcal{D}_m|^2\big|\mathcal{F}_{m-1}\big]\\ &\leq (C_\Psi)^2(\sigma_n)^{-2}\sum_{m=n_0(n)+1}^{n}\frac{b_m^2}{\bar b_n^2}\mathbb{E}\big[\mathbb{1}_{\{|D_m|>\frac{\varepsilon\bar b_n\sigma_n}{C_\Psi b_m}\}}|D_m|^2\big|\mathcal{F}_{m-1}\big]\to 0, \text{ in probability.}\end{aligned}$$

Thus, Theorem 5.1 implies that, on $\mathbb{U}_\infty^{\mathrm{conv}}$,

$$\frac{1}{\sigma_n}\Xi_n \overset{\text{stably}}{\Longrightarrow} A\,\mathcal{N}(0,\Gamma).$$

Together with step 2 (see Lemma B.4) we thus get that

$$\frac{1}{\sigma_n}\frac{1}{\bar b_n}\sum_{k=n_0(n)+1}^{n} b_k\,\bar{\mathcal{H}}_n[k,n]\,\pi_\theta(\Upsilon_k^{(n)}) \overset{\text{stably}}{\Longrightarrow} A\,\mathcal{N}(0,\Gamma), \text{ on } \mathbb{U}_\infty^{\mathrm{conv}}.$$

5) *Analysis of the contribution of* $\theta_{n_0(n)}$. By choice of $\mathbb{U}_n$, the asymptotic estimate (4.5) holds. This entails together with property (ii) of Lemma 3.6 that, on $\mathbb{U}_\infty$,

$$|\theta_{n_0(n)}| = d(X_{n_0(n)}, M) = \mathcal{O}_P(\sigma^{\mathrm{RM}}_{n_0(n)}).$$

Moreover, by Lemma 5.4, $\bar{\mathcal{H}}_{n_0(n)}[n_0(n),n]$ is uniformly bounded on $\mathbb{U}_\infty$, so that, on $\mathbb{U}_\infty$,

$$\frac{b_{n_0(n)}}{\bar b_n\gamma_{n_0(n)}}\bar{\mathcal{H}}_{n_0(n)}[n_0(n),n]\theta_{n_0(n)} = \mathcal{O}_P\Big(\frac{b_{n_0(n)}}{\bar b_n\gamma_{n_0(n)}}\sigma^{\mathrm{RM}}_{n_0(n)}\Big),$$

which is of order $o_P(\sigma_n)$ by assumption (2.12). With step 3 we thus obtain that, on $\mathbb{U}_\infty^{\mathrm{conv}}$,

$$\bar\theta_n \overset{\text{stably}}{\Longrightarrow} A\,\mathcal{N}(0,\Gamma).$$

6) *Comparison of* $\bar X_n$ *and* $\Phi(\bar\theta_n)$. On $\mathbb{U}_n^{\mathrm{conv}}$,

$$\begin{aligned}\bar X_n &= \frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\,\Phi(\zeta_m,\theta_m)\\ &= \frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\Big(\Phi(\bar\zeta_n,\bar\theta_n) + D\Phi(\bar\zeta_n,\bar\theta_n)\begin{pmatrix}\zeta_m-\bar\zeta_n\\ \theta_m-\bar\theta_n\end{pmatrix}\\ &\qquad + \mathcal{O}\big(|\zeta_m-\bar\zeta_n|^{1+\alpha_\Phi}+|\theta_m-\bar\theta_n|^{1+\alpha_\Phi}\big)\Big)\\ &= \Phi(\bar\zeta_n,\bar\theta_n) + \mathcal{O}\Big(\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m\big(|\zeta_m-\bar\zeta_n|^{1+\alpha_\Phi}+|\theta_m-\bar\theta_n|^{1+\alpha_\Phi}\big)\Big),\end{aligned}$$

where we used convexity of U_Φ and linearity of $D\Phi(\bar\zeta_n, \bar\theta_n)$. By Proposition 6.3, we get that

$$\sup_{m=n_0(n)+1,\dots,n} |\zeta_m - \zeta_{n_0(n)}| = \mathcal{O}_P\big(\varepsilon_n^{\mathrm{RM}}\big), \text{ on } \mathbb{U}_\infty^{\mathrm{conv}},$$

so that, on $\mathbb{U}_\infty^{\mathrm{conv}}$,

$$\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m|\zeta_m - \bar\zeta_n|^{1+\alpha_\Phi} \le \Big(2\sup_{m=n_0(n)+1,\dots,n}|\zeta_m - \zeta_{n_0(n)}|\Big)^{1+\alpha_\Phi} = \mathcal{O}_P\big((\varepsilon_n^{\mathrm{RM}})^{1+\alpha_\Phi}\big).$$

By assumption (2.13), the previous expression is of order $o_P(\sigma_n)$. Moreover, using that $|a-b|^{1+\alpha_\Phi} \le (|a|+|b|)^{1+\alpha_\Phi} \le 2^{\alpha_\Phi}(|a|^{1+\alpha_\Phi} + |b|^{1+\alpha_\phi})$ for $a, b \in \mathbb{R}^{d_\theta}$, $\sum_{m=n_0(n)+1}^n b_m = \bar b_n$ and Jensen's inequality we conclude that, on $\mathbb{U}_\infty$,

$$\begin{aligned}\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m|\theta_m - \bar\theta_n|^{1+\alpha_\Phi} &\le 2^{\alpha_\Phi}\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m(|\theta_m|^{1+\alpha_\Phi} + |\bar\theta_n|^{1+\alpha_\Phi}) \\ &\le 2^{1+\alpha_\Phi}\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m|\theta_m|^{1+\alpha_\Phi} \le 2^{1+\alpha_\Phi}\Big(\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m|\theta_m|^2\Big)^{(1+\alpha_\Phi)/2}.\end{aligned} \tag{7.3}$$

Recall that, on $\mathbb{U}_\infty$, $|\theta_m| = d(X_m, M)$ so that the bound of Theorem 4.1 implies that

$$\mathbb{E}\Big[\mathbb{1}_{\mathbb{U}_\infty}\frac{1}{\bar b_n}\sum_{m=n_0(n)}^{n} b_m|\theta_m|^2\Big] = \mathcal{O}\Big(\frac{1}{\bar b_n}\sum_{m=n_0(n)}^{n} b_m(\sigma_m^{\mathrm{RM}})^2\Big).$$

Using (7.3) we get, on $\mathbb{U}_\infty$,

$$\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m|\theta_m - \bar\theta_n|^{1+\alpha_\Phi} = \mathcal{O}_P\Big(\Big(\frac{1}{\bar b_n}\sum_{m=n_0(n)+1}^{n} b_m(\sigma_m^{\mathrm{RM}})^2\Big)^{(1+\alpha_\Phi)/2}\Big).$$

Hence, this term is of order $o_P(\sigma_n)$, on $\mathbb{U}_\infty$, by assumption (2.15). Altogether, we thus get that

$$\bar X_n = \Phi(\bar\zeta_n, \bar\theta_n) + o_P(\sigma_n), \text{ on } \mathbb{U}_\infty^{\mathrm{conv}}.$$

7) *Synthesis.* Note that on $\mathbb{U}_\infty^{\mathrm{conv}}$, from a random minimal n onwards all $\bar X_n$ lie in U_δ^ρ and Ψ is Lipschitz on U_δ^ρ, since it has regularity α_Ψ, so that we get with step 6 that

$$\Psi(\bar X_n) = \begin{pmatrix}\bar\zeta_n \\ \bar\theta_n\end{pmatrix} + o_P(\sigma_n), \text{ on } \mathbb{U}_\infty^{\mathrm{conv}}.$$

Consequently, by step 5, and Lemma B.4, one has

$$\sigma_n^{-1}\Psi_\theta(\bar X_n) \overset{\text{stably}}{\Longrightarrow} A\,\mathcal{N}(0,\Gamma), \text{ on } \mathbb{U}_\infty^{\mathrm{conv}}.$$

Now,

$$\sigma_n^{-1}(\Psi(\bar X_n) - \Psi(\bar X_n^*)) = \begin{pmatrix}0 \\ \sigma_n^{-1}\Psi_\theta(\bar X_n)\end{pmatrix} \overset{\text{stably}}{\Longrightarrow} \bar A\,\mathcal{N}(0,\Gamma), \text{ on } \mathbb{U}^{\mathrm{conv}},$$

with

$$\bar A = \begin{pmatrix}0 \\ A\end{pmatrix} = D\Psi(X_\infty)\big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1}\Pi_{N_{X_\infty}M}.$$

Here we used that the image of $Df(X_\infty)\big|_{N_{X_\infty}M}$ is in $N_{X_\infty}M$ so that

$$D\Psi_\zeta(X_\infty)\big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1}\Pi_{N_{X_\infty}M} = 0.$$

Next, note that, on $\mathbb{U}_\infty^{\mathrm{conv}}$,

$$\begin{aligned}\bar{X}_n - \bar{X}_n^* &= \Phi(\Psi(\bar{X}_n)) - \Phi(\Psi(\bar{X}_n^*)) \\ &= D\Phi(\Psi(\bar{X}_n^*))(\Psi(\bar{X}_n) - \Psi(\bar{X}_n^*)) + o(|\Psi(\bar{X}_n) - \Psi(\bar{X}_n^*)|)\end{aligned}$$

with $D\Phi(\Psi(\bar{X}_n^*)) \to D\Phi(\Psi(X_\infty))$, almost surely, on $\mathbb{U}_\infty^{\mathrm{conv}}$. Hence, $\sigma_n^{-1} D\Phi(\bar{X}_n^*)(\Psi(\bar{X}_n) - \Psi(\bar{X}_n^*))$ can be viewed as continuous function of $(D\Phi(\Psi(\bar{X}_n^*)), \sigma_n^{-1}(\Psi(\bar{X}_n) - \Psi(\bar{X}_n^*))$ which itself converges stably, on $\mathbb{U}_\infty^{\mathrm{conv}}$, by Lemma A.4. Moreover, the above error term is of order $o_P(\sigma_n^{-1})$, on $\mathbb{U}_\infty^{\mathrm{conv}}$, so that with Lemma B.4,

$$\sigma_n^{-1}(\bar{X}_n - \bar{X}_n^*) \overset{\text{stably}}{\Longrightarrow} Q\,\mathcal{N}(0,\Gamma), \text{ on } \mathbb{U}_\infty^{\mathrm{conv}},$$

with

$$Q = D\Phi(\Psi(X_\infty))\bar{A} = \big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1}\Pi_{N_{X_\infty}M} = B.$$

Thus, we proved (2.16).

Finally, on $\mathbb{U}_\infty^{\mathrm{conv}}$, for sufficiently large n Taylor together with the fact that $f(\bar{X}_n^*) = 0$ imply that

$$F(\bar{X}_n) - F(X_\infty) = \frac{1}{2} Df(\bar{X}_n^*)(\bar{X}_n - \bar{X}_n^*)^{\otimes 2} + o(|\bar{X}_n - \bar{X}_n^*|^2).$$

Moreover, using that $Df = D^2F$ is a symmetric matrix we conclude that

$$\begin{aligned}Df(\bar{X}_n^*)(\bar{X}_n - \bar{X}_n^*)^{\otimes 2} &= (\bar{X}_n - \bar{X}_n^*)^\dagger D^2F(\bar{X}_n^*)(\bar{X}_n - \bar{X}_n^*) \\ &= \big|(D^2F(\bar{X}_n^*))^{1/2}(\bar{X}_n - \bar{X}_n^*)\big|^2.\end{aligned}$$

Consequently, $\sigma_n^{-2} Df(\bar{X}_n^*)(\bar{X}_n - \bar{X}_n^*)^{\otimes 2}$ is a continuous function of $((D^2F(\bar{X}_n^*))^{1/2}$, $\sigma_n^{-1}(\bar{X}_n - \bar{X}_n^*))$ with the first component converging, almost surely, to $(D^2F(X_\infty))^{1/2}$, on $\mathbb{U}_\infty^{\mathrm{conv}}$, and the second component converging stably as derived above. Hence, we get stable convergence

$$2\sigma_n^{-2}(F(\bar{X}_n) - F(X_\infty)) \overset{\text{stably}}{\Longrightarrow} \big|\big(Df(X_\infty)\big|_{N_{X_\infty}M}\big)^{-1/2}\,\Pi_{N_{X_\infty}M}\,\mathcal{N}(0,\Gamma)\big|^2. \qquad \square$$

7.2 Proof of Theorem 2.6

Proof of Theorem 2.6. First, we verify that for every triple $(\alpha_f, \alpha_\Phi, \alpha_\Psi)$ as in (A.1) there exist γ and ρ satisfying (2.3) and that for every such γ and ρ there exists $(n_0(n))_{n\in\mathbb{N}}$ as in (A.3). By definition, $\alpha' > \frac{1}{2}$ so that every term on the left-hand side of γ in condition (2.3) is strictly smaller than one. Hence, γ and ρ can be chosen accordingly.

We prove existence of a $\mathbb{N}$-valued sequence $(n_0(n))$ with $0 \le n_0(n) < n$, $n_0(n) = o(n)$ and

$$n_0(n)^{-1} = o\Big(n^{-\frac{1}{2\gamma-1}\frac{1}{1+\alpha_\Phi}} \wedge n^{-\frac{1}{\alpha}\frac{1-\gamma}{2\gamma-1}}\Big).$$

By assumption (2.3), we have $\gamma > (1 - \frac{1}{2}\frac{\alpha_\Phi}{1+\alpha_\Phi}) \vee (1 - \frac{\alpha}{1+2\alpha})$ and elementary computations imply that

$$\frac{1}{2\gamma-1}\frac{1}{1+\alpha_\Phi} < 1 \text{ and } \frac{1}{\alpha}\frac{1-\gamma}{2\gamma-1} < 1.$$

Hence, the choice $n_0(n) = \lfloor n^\beta/2 \rfloor$ with $\lfloor\cdot\rfloor$ denoting the rounding off operation fulfils assumption (2.4) when choosing

$$\beta \in \Big(\frac{1}{2\gamma-1}\frac{1}{1+\alpha_\Phi} \vee \frac{1}{\alpha}\frac{1-\gamma}{2\gamma-1}, 1\Big).$$

Now, suppose that $\rho - \gamma < -1$. By assumption (2.3), we have

$$\gamma > \frac{1 + \frac{\alpha_\Phi}{2}}{1+\alpha_\Phi} > \frac{1}{1+\alpha_\Phi}, \tag{7.4}$$

so that we can additionally assume that $\beta > \big(\frac{1}{1+\alpha_\Phi} - (1+\rho)\big)/(\gamma - (1+\rho))$ since the right-hand side is strictly smaller than one. For this choice we then also have that

$$n_0(n)^{-1} = o\Big(n^{-\frac{\frac{1}{1+\alpha_\Phi}-(1+\rho)}{\gamma-(1+\rho)}}\Big).$$

Next, we verify the assumptions of Theorem 2.9 with $\sigma_n^{\mathrm{RM}} = n^{-\gamma/2}$ and $\delta_n^{\mathrm{diff}} \equiv 1$. Note that $\gamma > 1 - \frac{1}{2}\frac{\alpha_\Phi}{1+\alpha_\Phi}$ implies that $\gamma > \frac{3}{4}$.

(B.1) + (B.3): Immediate consequences of the assumptions.

(B.2): By definition of $(\gamma_n)_{n\in\mathbb{N}}$ one has $n\gamma_n \to \infty$ and $\gamma_n \to 0$. Furthermore, it is elementary to check that

$$\frac{b_{n+1}\gamma_n}{b_n\gamma_{n+1}} = 1 + (\rho+\gamma)n^{-1} + o(n^{-1}) = 1 + o(\gamma_n),$$

since $\gamma_n = n^{-\gamma}$ with $\gamma < 1$.

Moreover, note that

$$\frac{\sigma_{n-1}^{\mathrm{RM}} - \sigma_n^{\mathrm{RM}}}{\sigma_n^{\mathrm{RM}}} = \frac{\gamma}{2n} + o(n^{-1}) = o(\gamma_n)$$

and, trivially, $\sigma_{n-1}^{\mathrm{RM}} \approx \sigma_n^{\mathrm{RM}}$. By assumption (2.3), $2\rho > 2\gamma\alpha' - 2 > -1$. Hence,

$$\sum_{m=n_0(n)+1}^{n} (b_m\delta_m^{\mathrm{diff}})^2 \sim \int_{n_0(n)}^{n} s^{2\rho}\,\mathrm{d}s = \Big[\frac{1}{2\rho+1}s^{2\rho+1}\Big]_{n_0(n)}^{n} \sim \frac{1}{2\rho+1}n^{2\rho+1}.$$

Similarly, for $(L(n))$ as in (B.2)

$$\sum_{m=L(n)+1}^{n} (b_m\delta_m^{\mathrm{diff}})^2 \sim \int_{L(n)}^{n} s^{2\rho}\,\mathrm{d}s = \Big[\frac{1}{2\rho+1}s^{2\rho+1}\Big]_{L(n)}^{n} = o\big(n^{2\rho+1}\big),$$

since $L(n)^{2\rho+1} \sim n^{2\rho+1}$. Consequently,

$$\lim_{n\to\infty} \frac{\sum_{k=L(n)+1}^{n}(b_k\delta_k^{\mathrm{diff}})^2}{\sum_{k=n_0(n)+1}^{n}(b_k\delta_k^{\mathrm{diff}})^2} = 0.$$

(B.4): The almost sure convergence of $(\mathrm{cov}(D_m|\mathcal{F}_{m-1}))_{m\in\mathbb{N}}$ on $\mathbb{M}^{\mathrm{conv}}$ is true by assumption.

Let $x \in M$. According to (A.4), we can fix an open neighbourhood $U \subset \mathbb{R}^d$ of x such that $(\mathbb{1}_U(X_{n-1})|D_n|^2)_{n\in\mathbb{N}}$ is uniformly integrable and denote by $\mathbb{U}^{\mathrm{conv}}$ the event, that (X_n) converges to a point in $M \cap U$. Let $\varepsilon, \varepsilon' > 0$ arbitrary. To verify (2.10) we note that

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{(\sigma_n)^{-2}\sum_{m=n_0(n)+1}^{n}\frac{b_m^2}{\bar{b}_n^2}\mathbb{E}[\mathbb{1}_{\{|D_m|>\varepsilon\bar{b}_n\sigma_n/b_m\}}|D_m|^2|\mathcal{F}_{m-1}] > \varepsilon'\Big\} \cap \mathbb{U}^{\mathrm{conv}}\Big)\\
&\qquad \le \mathbb{P}\big(\big\{\exists m \in \{n_0(n)+1,\dots,n\} : X_{m-1} \notin U\big\} \cap \mathbb{U}^{\mathrm{conv}}\big)\\
&\qquad + \frac{1}{\varepsilon'}\mathbb{E}\Big[(\sigma_n)^{-2}\sum_{m=n_0(n)+1}^{n}\frac{b_m^2}{\bar{b}_n^2}\mathbb{E}[\mathbb{1}_U(X_{m-1})\mathbb{1}_{\{|D_m|>\varepsilon\bar{b}_n\sigma_n/b_m\}}|D_m|^2|\mathcal{F}_{m-1}]\Big]
\end{aligned}$$

and we will verify that the previous two summands converge to zero as $n \to \infty$.

The first term converges to zero, since on $\mathbb{U}^{\mathrm{conv}}$ the process stays in U from a random index onwards. To verify that also the second term tends to zero we observe that

$$\bar{b}_n = \sum_{m=n_0(n)+1}^{n} b_m \sim \frac{1}{\rho+1}n^{\rho+1},$$

so that

$$\sigma_n = \frac{1}{\bar{b}_n}\sqrt{\sum_{m=n_0(n)+1}^{n}(b_m\delta_m^{\mathrm{diff}})^2} \sim \frac{\rho+1}{\sqrt{2\rho+1}}n^{-1/2} \to 0 \tag{7.5}$$

and

$$\bar{b}_n\sigma_n \sim \frac{1}{\sqrt{2\rho+1}}n^{\rho+\frac{1}{2}} \text{ entails that } \inf_{m=n_0(n)+1,\dots,n}\bar{b}_n\sigma_n/b_m \to \infty \text{ as } n\to\infty,$$

since $\rho > -\frac{1}{2}$ and $b_m = m^\rho$. Hence, by the uniform integrability of $(\mathbb{1}_U(X_{m-1})|D_m|^2)_{m\in\mathbb{N}}$ we get that

$$\sup_{m=n_0(n)+1,\dots,n}\mathbb{E}[\mathbb{1}_U(X_{m-1})\mathbb{1}_{\{|D_m|>\varepsilon\bar{b}_n\sigma_n/b_m\}}|D_m|^2] \to 0 \tag{7.6}$$

and with $\sigma_n \to \infty$ we arrive at

$$\lim_{n\to\infty}(\sigma_n)^{-2}\sum_{m=n_0(n)+1}^{n}\frac{b_m^2}{\bar{b}_n^2}\mathbb{E}[\mathbb{1}_U(X_{m-1})\mathbb{1}_{\{|D_m|>\varepsilon\bar{b}_n\sigma_n/b_m\}}|D_m|^2] = 0,$$

so that we established convergence to zero in probability on $\mathbb{U}^{\mathrm{conv}}$. Similarly to 3.7, there exists a countable family $\mathcal{U}$ of open sets such that $(\mathbb{1}_U(X_{n-1})|D_n|^2)$ is uniformly integrable for all $U\in\mathcal{U}$ and

$$M \subset \bigcup_{U\in\mathcal{U}} U.$$

By the above argument, (2.10) holds on each $\mathbb{U}^{\mathrm{conv}}$ with $U\in\mathcal{U}$ and hence also on

$$\mathbb{M}^{\mathrm{conv}} = \bigcup_{U\in\mathcal{U}}\mathbb{U}^{\mathrm{conv}}.$$

Assumption (2.11) is true since $\sqrt{\gamma_n}/\sigma_n^{\mathrm{RM}} = \sqrt{C_\gamma}$ and $(\mathbb{E}[\mathbb{1}_U(X_{m-1})|D_m|^2])_{m\in\mathbb{N}}$ is uniformly bounded by uniform integrability.

The other assumptions of **(B.4)** are immediate consequences of **(A.4)** and the fact that $\delta_n^{\mathrm{diff}} \equiv 1$ and $\sigma_n^{\mathrm{RM}} = n^{-\gamma/2}$.

(B.5): Using that $\bar{b}_n \sim \frac{1}{\rho+1}n^{\rho+1}$ we conclude that

$$\sigma_n^{-1}\frac{b_{n_0(n)}}{\bar{b}_n\gamma_{n_0(n)}}\sigma_{n_0(n)}^{\mathrm{RM}} \sim \frac{\sqrt{2\rho+1}}{C_\gamma}n^{\frac{1}{2}-(\rho+1)}n_0(n)^{\rho+\gamma-\frac{\gamma}{2}} = \frac{\sqrt{2\rho+1}}{C_\gamma}\frac{n_0(n)^{\rho+\frac{\gamma}{2}}}{n^{\rho+\frac{1}{2}}}$$

which tends to zero since, by assumption (2.3), $\rho+\frac{1}{2} > \gamma\alpha' - \frac{1}{2} > 0$, $\gamma<1$ and $n_0(n)\le n$.

We verify that $(\varepsilon_n^{\mathrm{RM}})^{1+\alpha_\Phi} = o(\sigma_n)$.

$$\begin{aligned}
\varepsilon_n^{\mathrm{RM}} &= \sum_{m=n_0(n)+1}^{n}((\sqrt{\gamma_m}\sigma_m^{\mathrm{RM}})^{1+\alpha_\Psi} + \gamma_m(\sigma_{m-1}^{\mathrm{RM}})^{1+\alpha}) + \sqrt{\sum_{m=n_0(n)+1}^{n}\gamma_m(\sigma_m^{\mathrm{RM}})^2} \\
&\sim \sum_{m=n_0(n)+1}^{n}(C_\gamma^{\frac{1+\alpha_\Psi}{2}}m^{-\gamma(1+\alpha_\Psi)} + C_\gamma m^{-\gamma(1+\frac{1+\alpha}{2})}) + \sqrt{\sum_{m=n_0(n)+1}^{n}C_\gamma m^{-2\gamma}} \\
&= \mathcal{O}\Big(\sum_{m=n_0(n)+1}^{n}m^{-\gamma(1+\alpha')} + \sqrt{\sum_{m=n_0(n)+1}^{n}m^{-2\gamma}}\Big) \\
&= \mathcal{O}\Big(n_0(n)^{1-\gamma(1+\alpha')} + n_0(n)^{-\gamma+\frac{1}{2}}\Big),
\end{aligned}$$

where we used that $\gamma(1+\alpha')$ and 2γ are strictly bigger than 1 since $\gamma > \frac{3}{4}$ and $\alpha' > \frac{1}{2}$. By assumption $\gamma > \frac{1}{2\alpha'}$ so that $1-\gamma(1+\alpha') < -\gamma + \frac{1}{2}$ and $\varepsilon_n^{\mathrm{RM}} = \mathcal{O}\big(n_0(n)^{-\gamma+\frac{1}{2}}\big)$. By (2.4), we thus get that

$$(\varepsilon_n^{\mathrm{RM}})^{1+\alpha_\Phi} = \mathcal{O}\big(\big(n_0(n)^{-1}\big)^{(\gamma-\frac{1}{2})(1+\alpha_\Phi)}\big) = o\big(n^{-\frac{1}{2}}\big),$$

which is by (7.5) of order $o(\sigma_n)$.

We verify that $\varepsilon_n^{\mathrm{RP}} = o(\sigma_n)$. One has by definition of α'

$$\begin{aligned}
\varepsilon_n^{\mathrm{RP}} &= \frac{1}{\overline{b}_n} \sum_{m=n_0(n)+1}^{n} b_m\big(\gamma_m^{-\frac{1-\alpha_\Psi}{2}}(\sigma_m^{\mathrm{RM}})^{1+\alpha_\Psi} + (\sigma_{m-1}^{\mathrm{RM}})^{1+\alpha} + \sigma_{m-1}^{\mathrm{RM}}(\varepsilon_n^{\mathrm{RM}})^{\alpha}\big) \\
&= \mathcal{O}\Big(\frac{1}{n^{\rho+1}} \sum_{m=n_0(n)+1}^{n} m^{\rho}\big(\underbrace{m^{\gamma\frac{1-\alpha_\Psi}{2}-\frac{\gamma}{2}(1+\alpha_\Psi)}}_{=m^{-\gamma\alpha_\Psi}} + m^{-\gamma\frac{1+\alpha}{2}} + m^{-\frac{\gamma}{2}} n_0(n)^{-\alpha(\gamma-\frac{1}{2})}\big)\Big) \\
&= \mathcal{O}\Big(\frac{1}{n^{\rho+1}} \sum_{m=n_0(n)+1}^{n} m^{\rho}\big(m^{-\gamma\alpha'} + m^{-\frac{\gamma}{2}} n_0(n)^{-\alpha(\gamma-\frac{1}{2})}\big)\Big) \\
&= \mathcal{O}\big(n^{-\gamma\alpha'} + n^{-\frac{\gamma}{2}} n_0(n)^{-\alpha(\gamma-\frac{1}{2})}\big),
\end{aligned}$$

where we used that $\rho - \gamma\alpha' > -1$ and $\rho - \frac{\gamma}{2} > -1$ as consequence of (2.3). Recall that by assumption $\gamma\alpha' > \frac{1}{2}$ and $n_0(n)^{-1} = o(n^{-\frac{1}{\alpha}\frac{1-\gamma}{2\gamma-1}})$ so that $\varepsilon_n^{\mathrm{RP}} = o(n^{-\frac{1}{2}}) = o(\sigma_n)$.

Finally, we show that

$$\frac{1}{\overline{b}_n} \sum_{m=n_0(n)+1}^{n} b_m(\sigma_m^{\mathrm{RM}})^2 = o(n^{-\frac{1}{1+\alpha_\Phi}}).$$

We have

$$\frac{1}{\overline{b}_n} \sum_{m=n_0(n)+1}^{n} b_m(\sigma_m^{\mathrm{RM}})^2 \sim \frac{\rho+1}{n^{\rho+1}} \sum_{m=n_0(n)+1}^{n} n^{\rho-\gamma},$$

so that in the case where $\rho - \gamma > -1$ the latter term is of order $O(n^{-\gamma}) = o(n^{-\frac{1}{1+\alpha_\Phi}})$ as consequence of (7.4). In the case where $\rho - \gamma = 1$ we use that $\rho + 1 = \gamma > 1/(1+\alpha_\Phi)$ to conclude that

$$\frac{1}{n^{\rho+1}} \sum_{m=n_0(n)+1}^{n} m^{-1} \le \frac{1}{n^{\rho+1}} \log(n) = o(n^{-\frac{1}{1+\alpha_\Phi}}).$$

Finally, in the case where $\rho - \gamma < -1$ with (2.5)

$$\frac{1}{n^{\rho+1}} \sum_{m=n_0(n)+1}^{n} m^{\rho-\gamma} = \mathcal{O}\Big(\frac{n_0(n)^{-\gamma+\rho+1}}{n^{\rho+1}}\Big) = o(n^{-\frac{1}{1+\alpha_\Phi}}). \qquad \square$$

A Stable convergence

In this section, we introduce the concept of stable convergence on a set. It is a slight generalisation of stable convergence introduced in [32].

Definition A.1. *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^d$-valued random variables, $A \in \mathcal{F}$ and K a probability kernel from $(A, \mathcal{F}|_A)$ to $(\mathbb{R}^d, \mathcal{B}^d)$. We say that (Y_n)* converges stably on A to K *and write*

$$Y_n \overset{\text{stably}}{\Longrightarrow} K, \quad \text{on } A,$$

if for every $B \in \mathcal{F}$ and continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_{A\cap B} f(Y_n)\big] = \mathbb{E}\Big[\mathbb{1}_{A\cap B} \int f(y)\, K(\cdot, dy)\Big]. \tag{A.1}$$

In the case where $A = \Omega$, we briefly say that (Y_n) converges stably to K *and write*

$$Y_n \overset{\text{stably}}{\Longrightarrow} K.$$

We give some central properties of stable convergence.

Theorem A.2. *Let (Y_n), A and K as in the previous definition and let $\mathcal{E}$ denote a $\cap$-stable generator of $\mathcal{F}$ containing Ω. The following properties are equivalent.*

(i) (Y_n) *converges stably to K on A.*

(ii) For every $B \in \mathcal{E}$ and continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_{A\cap B} f(Y_n)\big] = \mathbb{E}\Big[\mathbb{1}_{A\cap B} \int f(y)\, K(\cdot, dy)\Big].$$

(iii) For every $B \in \mathcal{E}$ and $\xi \in \mathbb{R}^d$

$$\lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_{A\cap B} e^{i\langle \xi, (Y_n)\rangle}\big] = \mathbb{E}\Big[\mathbb{1}_{A\cap B} \int e^{i\langle \xi, y\rangle}\, K(\cdot, dy)\Big].$$

(iv) For every bounded random variable Υ and every bounded and continuous $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_A \Upsilon f(Y_n)\big] = \mathbb{E}\Big[\mathbb{1}_A \Upsilon \int f(y)\, K(\cdot, dy)\Big].$$

Proof. (**ii**) $\Rightarrow$ (**i**) : First, suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is non-negative. It is standard to verify that the set $\mathcal{F}^f$ of all sets $B \in \mathcal{F}$ with the property that

$$\lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_{A\cap B} f(Y_n)\big] = \mathbb{E}\Big[\mathbb{1}_{A\cap B} \int f(y)\, K(\cdot, dy)\Big]$$

is a Dynkin-system. Since $\mathcal{F}^f$ contains the generator $\mathcal{E}$ we thus have $\mathcal{F}^f = \mathcal{F}$ and we verified property (A.1) for non-negative $f : \mathbb{R}^d \to \mathbb{R}$. For a general bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$ we write $f = \bar{f} - c$ with a non-negative function $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ and a constant $c \geq 0$. Clearly, (A.1) holds for $\bar{f}$ and the constant function c and by linearity of the integral and the limit we get that (A.1) also holds for $f = \bar{f} - c$.

(**iii**) $\Rightarrow$ (**ii**) : Follows from [20, Cor 3.8] where we set in the notation of the corollary $\mathcal{G} = \mathcal{F}|_A$ with the $\cap$-stable generator $\{A \cap B | B \in \mathcal{E}\}$.

(**i**) $\Rightarrow$ (**iv**) : For non-negative f and Υ, the asymptotic property follows by a monotone class argument and the general case is derived by using linearity. □

Lemma A.3. *(i) Let $A, A' \in \mathcal{F}$ and suppose that (Y_n) converges stably to K and K' on A and A', respectively. Then for almost all $\omega \in A \cap A'$ one has*

$$K(\omega, \cdot) = K'(\omega, \cdot).$$

In particular, the kernel appearing as limit is unique up to almost sure equivalence.

(ii) Let $(A_m)_{m\in\mathbb{N}}$ be a subfamily of $\mathcal{F}$ and suppose that for each $m \in \mathbb{N}$, (Y_n) converges stably to K_m on A_m. Then there exists a probability kernel K from $A := \bigcup_{m\in\mathbb{N}} A_m$ to $\mathbb{R}^d$ such that for all $m \in \mathbb{N}$ and almost all $\omega \in A_m$

$$K(\omega, \cdot) = K_m(\omega, \cdot)$$

and for every such kernel K we have

$$Y_n \overset{\text{stably}}{\Longrightarrow} K, \text{ on } A.$$

Proof. **(i):** We first show uniqueness of stable limits. By basic measure theory, there exists a countable set of bounded and continuous functions $f_n : \mathbb{R}^d \to \mathbb{R}$ $(n \in \mathbb{N})$ that characterize a probability distribution on $\mathbb{R}^d$. That means for two distributions μ and μ' on $\mathbb{R}^d$ one has the equivalence

$$\mu = \mu' \iff \forall n \in \mathbb{N} : \int f_n \, d\mu = \int f_n \, d\mu'.$$

Suppose now that (Y_n) converges to K and K' on a set $A \in \mathcal{F}$. Let $n \in \mathbb{N}$ and

$$B_n^+ = \Big\{\omega \in A : \int f_n(y) \, K(\omega, dy) > \int f_n(y) \, K'(\omega, dy)\Big\}.$$

Then,

$$\mathbb{E}\Big[\mathbb{1}_{B_n^+} \int f_n(y) \, K(\cdot, dy)\Big] \leftarrow \mathbb{E}\big[\mathbb{1}_{B_n^+} f_n(Y_m)\big] \to \mathbb{E}\Big[\mathbb{1}_{B_n^+} \int f_n(y) \, K'(\cdot, dy)\Big],$$

so that

$$\mathbb{E}\Big[\mathbb{1}_{B_n^+}\Big(\int f_n(y) \, K(\cdot, dy) - \int f_n(y) \, K'(\cdot, dy)\Big)\Big] = 0$$

and B_n^+ is a nullset. With the same argument we obtain that the event defined as B_n^+ with $>$ replaced by $<$, say B_n^- is a nullset. Consequently, $B = \bigcup B_n^+ \cup \bigcup B_n^-$, is a nullset and for every $\omega \in A \backslash B$ we have $K(\omega, \cdot) = K'(\omega, \cdot)$ due to the choice of $(f_n : n \in \mathbb{N})$.

Now suppose that K and K' are the stable limits of (Y_n) on two distinct sets A and A', respectively. As one easily verifies the restrictions of K and K' to $A \cap A'$ are stable limits of (Y_n) on $A \cap A'$ and thus they agree by the first part up to almost sure equivalence.

(ii): We first define a kernel K and verify that it is the stable limit on A. Note that $A'_m := A_m \backslash \bigcup_{k=1}^{m-1} A_k$ defines a partition $(A'_m)_{m \in \mathbb{N}}$ of A and set for $\omega \in A$

$$K(\omega, \cdot) = \sum_{m \in \mathbb{N}} \mathbb{1}_{A'_m} K_m(\omega, \cdot). \tag{A.2}$$

Fix $B \in \mathcal{F}$ and a bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$. We set $B_m = B \backslash \bigcup_{k=1}^{m-1} A_k$ and use stable convergence to K_m on A_m to conclude that

$$\begin{aligned}\mathbb{E}\big[\mathbb{1}_{A'_m \cap B} f(Y_n)\big] = \mathbb{E}\big[\mathbb{1}_{A_m \cap B_m} f(Y_n)\big] &\to \mathbb{E}\Big[\mathbb{1}_{A_m \cap B_m} \int f(y) \, K_m(\cdot, dy)\Big] \\ &= \mathbb{E}\Big[\mathbb{1}_{A'_m \cap B} \int f(y) \, K_m(\cdot, dy)\Big].\end{aligned}$$

Now dominated convergence implies that

$$\begin{aligned}\mathbb{E}\big[\mathbb{1}_{A \cap B} f(Y_n)\big] = \sum_{m \in \mathbb{N}} \mathbb{E}\big[\mathbb{1}_{A'_m \cap B} f(Y_n)\big] &\to \sum_{m \in \mathbb{N}} \mathbb{E}\Big[\mathbb{1}_{A'_m \cap B} \int f(y) \, K_m(\cdot, dy)\Big] \\ &= \mathbb{E}\Big[\mathbb{1}_{A \cap B} \int f(y) \, K(\cdot, dy)\Big],\end{aligned}$$

where the integrable majorant is given by $(C \, \mathbb{P}(A'_m))_{m \in \mathbb{N}}$ with $C > 0$ being a uniform bound for f. We thus showed stable convergence on A to the particular kernel K. Note that the previous arguments also apply for any kernel K with the property that for all $m \in \mathbb{N}$ and almost all $\omega \in A_m$, $K(\omega, \cdot) = K_m(\omega, \cdot)$. It thus remains to show that the particular kernel possesses the latter property. However, this is an immediate consequence of part (i) since (Y_n) converges stably to $K|_{A_m}$ on A_m so that $K|_{A_m}$ and K_m agree up to nullsets. □

Lemma A.4. *Let $d' \in \mathbb{N}$ and $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^{d'}$-valued random variables that converges, in probability, on A, to a $\mathbb{R}^{d'}$-valued random variable X_∞. If $(Y_n)_{n\in\mathbb{N}}$ converges stably to K on A, then the extended sequence $(X_n, Y_n)_{n\in\mathbb{N}}$ converges stably, on A to the kernel*

$$\bar{K}(\omega, d(x,y)) = \delta_{X_\infty(\omega)}(dx)\, K(\omega, dy).$$

Proof. Choosing $\mathcal{G} = \mathcal{F}|_A$, $Y = X_\infty \mathbb{1}_A$, $Y_n = X_n \mathbb{1}_A$ and $(X_n) = (Y_n)$ in Theorem 3.7 of [20] yields

$$(\mathbb{1}_A X_n, Y_n) \overset{\text{stably}}{\Longrightarrow} \delta_{\mathbb{1}_A X_\infty} \otimes K, \text{ on } A,$$

so that for every $B \in \mathcal{F}$ and continuous and bounded function $f : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_{A\cap B} f(X_n, Y_n)\big] = \mathbb{E}\Big[\mathbb{1}_{A\cap B} \int f(x,y)\, \delta_{X_\infty}(dx) K(\cdot, dy)\Big]. \qquad \square$$

We will use a classical central limit theorem for martingales, see [19]. A consequence of [19, Corollary 3.1] is the following theorem. In contrast to the original version the statement allows multidimensional processes. However, this generalisation is easily obtained by noticing that it suffices to prove the central limit theorem for linear functionals of the process.

Theorem A.5. *For every $n \in \mathbb{N}$ let $(Z_i^{(n)})_{i=1,\dots,k_n}$ be a sequence of $\mathbb{R}^d$-valued martingale differences for a filtration $(\mathcal{F}_i^{(n)})_{i=1,\dots,k_n}$ with $\mathcal{F}_i^{(n)} \subset \mathcal{F}_i^{(n+1)}$ for all $i = 1, \dots, k_n$. Suppose that the following holds:*

(i) $\forall \varepsilon > 0 : \sum_{i=1}^{k_n} \mathbb{E}\big[\mathbb{1}\{|Z_i^{(n)}| > \varepsilon\}\, |Z_i^{(n)}|^2 \big| \mathcal{F}_{i-1}^{(n)}\big] \to 0$, *in probability, and*

(ii) $\sum_{i=1}^{k_n} \operatorname{cov}(Z_i^{(n)} | \mathcal{F}_{i-1}^{(n)}) \to \Gamma$, *in probability.*

Then

$$\sum_{i=1}^{k_n} Z_i^{(n)} \overset{\text{stably}}{\Longrightarrow} \mathcal{N}(0, \Gamma).$$

We extend the theorem to restricted stable convergence.

Theorem A.6. *For every $n \in \mathbb{N}$, let $(Z_i^{(n)})_{i=1,\dots,k_n}$ be a sequence of $\mathbb{R}^d$-valued martingale differences for a fixed filtration $(\mathcal{F}_i)_{i\in\mathbb{N}}$ and let $A \in \mathcal{F}_\infty = \bigvee_{i\in\mathbb{N}} \mathcal{F}_i$. Suppose that $\lim_{n\to\infty} k_n = \infty$ and the following holds:*

(i) $\forall \varepsilon > 0 : \sum_{i=1}^{k_n} \mathbb{E}\big[\mathbb{1}\{|Z_i^{(n)}| > \varepsilon\}\, |Z_i^{(n)}|^2 \big| \mathcal{F}_{i-1}\big] \to 0$, *in probability, on A, and*

(ii) $\sum_{i=1}^{k_n} \operatorname{cov}(Z_i^{(n)} | \mathcal{F}_{i-1}) \to \Gamma$, *in probability, on A.*

Then

$$\sum_{i=1}^{k_n} Z_i^{(n)} \overset{\text{stably}}{\Longrightarrow} \mathcal{N}(0, \Gamma), \quad \text{on } A.$$

Remark A.7. In the theorem one can replace assumption (i) by the stronger assumption that there exists $q > 2$ with

$$\sum_{i=1}^{k_n} \mathbb{E}\big[|Z_i^{(n)}|^q \big| \mathcal{F}_{i-1}^{(n)}\big] \to 0, \text{ in probability, on } A.$$

Indeed, this follows since $\mathbb{1}\{|Z_i^{(n)}| > \varepsilon\}\, |Z_i^{(n)}|^2 \le \varepsilon^{-(q-2)} |Z_i^{(n)}|^q$.

Proof of Theorem A.6. Applying a diagonalisation argument on property (i) we deduce existence of two zero sequences $(\delta_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ of positive reals with

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big\{\sum_{i=1}^{k_n} \mathbb{E}[\mathbb{1}_{\{|Z_i^{(n)}|>\varepsilon_n\}}|Z_i^{(n)}|^2|\mathcal{F}_{i-1}] > \delta_n\Big\} \cap A\Big) = 0.$$

We fix $\delta \in (0,1)$ and set $I_n = \mathbb{E}[\mathbb{1}_A|\mathcal{F}_n]$ for all $n \in \mathbb{N}$ and consider the stopping times

$$T^{(n)} = \inf\Big\{m = 0,\dots,k_n-1 : I_m \le \delta \text{ or } \sum_{i=1}^{m+1} \mathbb{E}[\mathbb{1}_{\{|Z_i^{(n)}|>\varepsilon_n\}}|Z_i^{(n)}|^2|\mathcal{F}_{i-1}] > \delta_n\Big\}$$

with the infimum of the empty set being ∞. We will apply Theorem A.5 onto $(\bar{Z}_i^{(n)})_{i=1,\dots,k_n}$ given by

$$\bar{Z}_i^{(n)} = \mathbb{1}_{\{T^{(n)}\ge i\}} Z_i^{(n)}.$$

We verify assumptions (i) and (ii). First note that for every $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$, $\varepsilon_n \le \varepsilon$ and for those n we get that

$$\begin{aligned}\sum_{i=1}^{k_n} \mathbb{E}\big[\mathbb{1}\{|\bar{Z}_i^{(n)}| > \varepsilon\}\,|\bar{Z}_i^{(n)}|^2\big|\mathcal{F}_{i-1}\big] &\le \sum_{i=1}^{k_n} \mathbb{E}\big[\mathbb{1}\{|\bar{Z}_i^{(n)}| > \varepsilon_n\}\,|\bar{Z}_i^{(n)}|^2\big|\mathcal{F}_{i-1}\big] \\ &= \sum_{i=1}^{k_n} \mathbb{1}_{\{T^{(n)}\ge i\}}\mathbb{E}\big[\mathbb{1}\{|Z_i^{(n)}| > \varepsilon_n\}\,|Z_i^{(n)}|^2\big|\mathcal{F}_{i-1}\big] \le \delta_n \to 0.\end{aligned}$$

Second, $(I_n)_{n\in\mathbb{N}}$ is a martingale that converges to $\mathbb{E}[\mathbb{1}_A|\mathcal{F}_\infty] = \mathbb{1}_A$, a.s., so that up to nullsets $A^{(\delta)} := \{\min_{n\in\mathbb{N}} I_n > \delta\} \subset A$. Furthermore, $\mathbb{P}(A^{(\delta)}\Delta\{T^{(n)} = \infty\}) \to 0$ as $n \to \infty$. Thus we have, with high probability, on $A^{(\delta)}$,

$$\sum_{i=1}^{k_n} \mathrm{cov}(\bar{Z}_i^{(n)}|\mathcal{F}_{i-1}^{(n)}) = \sum_{i=1}^{k_n} \mathbb{1}_{\{T^{(n)}\ge i\}}\mathrm{cov}(Z_i^{(n)}|\mathcal{F}_{i-1}) = \sum_{i=1}^{k_n} \mathrm{cov}(Z_i^{(n)}|\mathcal{F}_{i-1}) \to \Gamma.$$

Conversely, on $(A^{(\delta)})^c$ the stopping time $T = \inf\{m \in \mathbb{N} : I_m \le \delta\}$ is finite and we get on $(A^{(\delta)})^c$

$$\begin{aligned}\sum_{i=1}^{k_n} \|\mathrm{cov}(\bar{Z}_i^{(n)}|\mathcal{F}_{i-1})\| &\le \sum_{i=1}^{k_n} \mathbb{1}_{\{T^{(n)}\ge i\}}\mathbb{E}[|Z_i^{(n)}|^2|\mathcal{F}_{i-1}] \le \sum_{i=1}^{k_n} \mathbb{1}_{\{T^{(n)}\ge i\}}\mathbb{E}\big[|Z_i^{(n)}|^2\big|\mathcal{F}_{i-1}\big] \\ &\le \sum_{i=1}^{k_n} \mathbb{1}_{\{T^{(n)}\ge i\}}\big(\mathbb{E}\big[\mathbb{1}\{|Z_i^{(n)}| > \varepsilon_n\}|Z_i^{(n)}|^2\big|\mathcal{F}_{i-1}\big] + \varepsilon_n\big) \\ &\le \big(\delta_n + T\varepsilon_n\big) \to 0.\end{aligned}$$

Thus, we showed that

$$\sum_{i=1}^{k_n} \bar{Z}_i^{(n)} \overset{\text{stably}}{\Longrightarrow} \mathcal{N}(0, \mathbb{1}_{A^{(\delta)}}\Gamma).$$

Recalling that on $A^{(\delta)}$, with high probability, $\sum_{i=1}^{k_n} Z_i^{(n)} = \sum_{i=1}^{k_n} \bar{Z}_i^{(n)}$ we conclude that

$$\sum_{i=1}^{k_n} Z_i^{(n)} \overset{\text{stably}}{\Longrightarrow} \mathcal{N}(0,\Gamma), \quad \text{on } A^{(\delta)}.$$

Finally, we note that (I_n) takes values in $[0,1]$ and once the process hits zero it stays there, almost surely. Hence, one has $A = \{\min_{n\in\mathbb{N}} I_n > 0\}$ up to nullsets. This implies that up to nullsets

$$A = \bigcup_{\delta>0} A^{(\delta)}$$

Thus, an application of Lemma A.3 finishes the proof. □

B $\mathcal{O}_P$ and o_P

We will use the $\mathcal{O}$- and o-notation in a probabilistic sense.

Definition B.1. *Let $A \in \mathcal{F}$, (X_n) be a sequence of $\mathbb{R}^d$-valued random variables and (a_n) be a sequence of strictly positive reals.*

(1) If

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(\{|X_n| > Ca_n\} \cap A) = 0,$$

we say that (X_n) is of order $\mathcal{O}(a_n)$, in probability, on A, *and write*

$$X_n = \mathcal{O}_P(a_n), \text{ on } A.$$

(2) If for every $C > 0$

$$\limsup_{n\to\infty} \mathbb{P}(\{|X_n| > Ca_n\} \cap A) = 0,$$

we say that (X_n) is of order $o(a_n)$, in probability, on A, *and write*

$$X_n = o_P(a_n), \text{ on } A.$$

Remark B.2. Expectations together with Markov's inequality are an efficient tool for verifying that a sequence (X_n) of random variables is of order $\mathcal{O}(a_n)$. Indeed,

$$\limsup_{n\to\infty} \mathbb{P}(\{|X_n| > Ca_n\} \cap A) \le \frac{1}{C} \limsup_{n\to\infty} \frac{\mathbb{E}[\mathbb{1}_A |X_n|]}{a_n},$$

so that finiteness of the $\limsup$ on the right implies that $X_n = \mathcal{O}_P(a_n)$, on A.

Lemma B.3. *Let (a_n) be a sequence of strictly positive reals, (X_n) be a sequence of $\mathbb{R}^d$-valued random variables and $A, A_1, A_2, \ldots \in \mathcal{F}$ with $\mathbb{P}(A \setminus \bigcup_{m\in\mathbb{N}} A_m) = 0$. If for every $m \in \mathbb{N}$*

$$X_n = \mathcal{O}_P(a_n), \quad \text{on } A_m,$$

then

$$X_n = \mathcal{O}_P(a_n), \text{ on } A.$$

Proof. Let $\varepsilon > 0$ and choose $M \in \mathbb{N}$ such that $\mathbb{P}(A \setminus \bigcup_{m=1}^M A_m) \le \varepsilon$. Now,

$$\mathbb{P}(\{X_n \ge Ca_n\} \cap A) \le \sum_{m=1}^M \mathbb{P}(\{X_n \ge Ca_n\} \cap A_m) + \mathbb{P}\Big(A \setminus \bigcup_{m=1}^M A_m\Big),$$

so that

$$\limsup_{n\to\infty} \mathbb{P}(\{X_n \ge Ca_n\} \cap A) \le \sum_{m=1}^M \limsup_{n\to\infty} \mathbb{P}(\{X_n \ge Ca_n\} \cap A_m) + \varepsilon.$$

Consequently,

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(\{X_n \ge Ca_n\} \cap A) \le \varepsilon$$

and the statement follows since $\varepsilon > 0$ was arbitrary. □

Lemma B.4. *Let $A \in \mathcal{F}$ and $(X_n), (Y_n)$ be $\mathbb{R}^d$-valued sequences of random variables. Suppose that (Y_n) converges stably to K on A and $X_n = o_P(1)$, on A. Then*

$$X_n + Y_n \overset{\text{stably}}{\Longrightarrow} K, \text{ on } A.$$

Proof. Let $\varepsilon > 0$. By the assumptions on (X_n), we have

$$\limsup_{n\to\infty} \mathbb{P}(\{|X_n| > \varepsilon\} \cap A) = 0,$$

so that

$$X_n \to 0, \text{ in probability, on } A.$$

Thus, with Lemma A.4,

$$(X_n, Y_n) \overset{\text{stably}}{\Longrightarrow} \delta_0 \otimes K, \text{ on } A.$$

Define

$$g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d; \quad (x, y) \mapsto x + y.$$

Let $B \in \mathcal{F}$ and $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded. Then,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{A\cap B} f(X_n + Y_n)] = \mathbb{E}[\mathbb{1}_{A\cap B}(f \circ g)(X_n, Y_n)] &\to \mathbb{E}\left[\mathbb{1}_{A\cap B} \iint f(x+y)\ \delta_0(dx) K(\cdot, dy)\right] \\ &= \mathbb{E}\left[\mathbb{1}_{A\cap B} \int f(y)\ K(\cdot, dy)\right]. \qquad \square\end{aligned}$$

C Nice representations in the sense of Def. 2.4, Fermi coordinates

In this section we discuss the existence of nice representations.

Lemma C.1. *Let $d_\zeta \in \{1, \dots, d-1\}$ and $M \subset \mathbb{R}^d$ be a d_ζ-dimensional C^3-submanifold. Then every $x \in M$ admits a nice representation $\Phi : U_\Phi \to U$ for a neighbourhood U of x that is C^2.*

Proof. We use Fermi coordinates. Let U be an open neighbourhood of x and $\Gamma : U_\Gamma \to U$ a C^3-diffeomorphism with

$$\Gamma(M_\Gamma \times \{0\}^{d_\theta}) = U \cap M, \quad \text{where } M_\Gamma := \{\zeta \in \mathbb{R}^{d_\zeta} : (\zeta, 0) \in U_\Gamma\}$$

and $d_\theta = d - d_\zeta$. We define a mapping

$$\tilde{\Phi} : M_\Gamma \times \mathbb{R}^{d_\theta} \to \mathbb{R}^d$$

as follows. For every $\zeta \in M_\Gamma$ we apply the Gram-Schmidt orthonormalisation procedure to the column vectors of the invertible matrix $D\Gamma(\zeta, 0)$ that is the vectors $D\Gamma(\zeta,0)e_1, \dots, D\Gamma(\zeta,0)e_d$ with $e_1, \dots, e_d$ denote the standard basis of $\mathbb{R}^d$. That means we iteratively set for $k = 1, \dots, d$

$$\bar{e}_k(\zeta) = \frac{D\Gamma(\zeta,0)e_k - \sum_{i=1}^{k-1} \langle \bar{e}_i(\zeta), D\Gamma(\zeta,0)e_k\rangle\, \bar{e}_i(\zeta)}{|D\Gamma(\zeta,0)e_k - \sum_{i=1}^{k-1} \langle \bar{e}_i(\zeta), D\Gamma(\zeta,0)e_k\rangle\, \bar{e}_i(\zeta)|}.$$

By induction over k it easily follows that the mapping $\zeta \mapsto \bar{e}_k(\zeta)$ is C^2 and we set

$$\tilde{\Phi} : M_\Gamma \times \mathbb{R}^{d_\theta} \to \mathbb{R}^d,\ (\zeta, \theta) \mapsto \Gamma(\zeta, 0) + \sum_{i=1}^{d_\theta} \theta_i \bar{e}_{d_\zeta + i}(\zeta).$$

Note that $\tilde{\Phi}$ is C^2 and $\bar{e}_{d_\zeta+1}(\zeta), \dots, \bar{e}_d(\zeta)$ span the normal space $N_{\Gamma(\zeta,0)}M$. We differentiate $\tilde{\Phi}$ in $(\zeta, 0)$ with $\zeta \in M_\Gamma$. One has for every $k = 1, \dots, d_\zeta$ and $\ell = 1, \dots, d_\theta$,

$$\frac{\partial}{\partial \zeta_k}\tilde{\Phi}(\zeta, 0) = \frac{\partial}{\partial \zeta_k}\Gamma(\zeta, 0) \text{ and } \frac{\partial}{\partial \theta_\ell}\tilde{\Phi}(\zeta, 0) = \bar{e}_{d_\zeta+\ell}(\zeta).$$

By construction, the first d_ζ columns of $D\tilde{\Phi}(\zeta,0)$ are linearly independent and span the same linear space as the vectors $\bar{e}_1(\zeta),\dots,\bar{e}_{d_\zeta}(\zeta)$ so that all columns of $D\tilde{\Phi}(\zeta,0)$ are linearly independent and $D\tilde{\Phi}(\zeta,0)$ is an invertible matrix. We set $(\zeta_0,0)=\Gamma^{-1}(x)$ and note that the mapping $\tilde{\Phi}$ restricted to an appropriate ball $B_{r_0}(\zeta_0,0)\subset M_\Gamma\times\mathbb{R}^{d_\theta}$ is a C^2-diffeomorphism onto its image.

Possibly, $\big(\tilde{\Phi}|_{B_{r_0}(\zeta_0,0)}\big)^{-1}(M)$ is not a subset of $\mathbb{R}^{d_\zeta}\times\{0\}^{d_\theta}$. Since the manifold M has no boundary we can choose $r_1\in(0,r_0)$ such that $K:=\tilde{\Phi}\big(\overline{B_{r_0}(\zeta_0,0)}\big)\cap M$ is compact. Hence, there exists $r_2\in(0,r_1)$ such that for all $x\in K$ and $y\in N_xM$ with $|y|\le r_2$, x is the unique closest element to $x+y$ in M, see [11, Theorem 3.2]. In particular, $x+y\notin M$ if $y\neq 0$. Consequently, for $(\zeta,\theta)\in B_{r_2}(\zeta_0,0)$ with $\theta\neq 0$ we have

$$\tilde{\Phi}(\zeta,\theta)\notin M,$$

so that $\big(\tilde{\Phi}|_{B_{r_2}(\zeta_0,0)}\big)^{-1}(M)\subset\mathbb{R}^{d_\zeta}\times\{0\}^{d_\theta}$. Altogether, we thus proved that the restriction of $\tilde{\Phi}|_{B_{r_2}(\zeta_0,0)}$ is a nice representation for M on $\tilde{\Phi}(B_{r_2}(\zeta_0,0))\ni x$. □

For a general introduction into Fermi coordinates of Riemannian submanifolds we refer the reader to chapter 2 of [18].

D Locality of the Robbins-Monro scheme

In this section, we prove a locality result for the Robbins-Monro scheme $\mathbf{X}=(X_n)_{n\in\mathbb{N}_0}$ defined in (1.1).

Lemma D.1. *Suppose that $F:\mathbb{R}^d\to\mathbb{R}$ is a C^1-function with Lipschitz continuous differential f that satisfies*

$$\lim_{|x|\to\infty}F(x)=-\infty.$$

Moreover, suppose that $(D_n)_{n\in\mathbb{N}}$ is a sequence of martingale differences that satisfies for a $C_D\ge 0$ and a bounded sequence of strictly positive reals $(\sigma_n^{\mathrm{RM}})_{n\in\mathbb{N}}$ that, almost surely,

$$\big(\frac{\sigma_n^{\mathrm{RM}}}{\sqrt{\gamma_n}}\big)^{-1}\,\mathbb{E}[|D_n|^2|\mathcal{F}_{n-1}]^{1/2}\le C_D, \tag{D.1}$$

for all but finitely many $n\in\mathbb{N}$. Moreover, assume that $\gamma_n\to 0$ and $\sum_{n=1}^\infty\gamma_n(\sigma_n^{\mathrm{RM}})^2<\infty$. Then

$$\mathbb{P}\Big(\limsup_{n\to\infty}|X_n|<\infty\Big)=1.$$

Proof. It suffices to show that $\mathbb{P}(\liminf_{n\to\infty}F(X_n)>-\infty)=1$. Let $x,y\in\mathbb{R}^d$ and note that with the Lipschitz continuity of f we get

$$\begin{aligned}F(y)&=F(x)+\int_0^1\langle f(x+t(y-x)),y-x\rangle\,dt\\&=F(x)+\langle f(x),y-x\rangle+\int_0^1\langle f(x+t(y-x))-f(x),y-x\rangle\,dt\\&\ge F(x)+\langle f(x),y-x\rangle-\frac{1}{2}\|f\|_{\mathrm{Lip}(\mathbb{R}^d)}\,|y-x|^2.\end{aligned}$$

Applying this to $x=X_{n-1}$ and $y=X_n$ gives

$$F(X_n)-F(X_{n-1})\ge\gamma_n|f(X_{n-1})|^2+\gamma_n\langle f(X_{n-1}),D_n\rangle-\|f\|_{\mathrm{Lip}(\mathbb{R}^d)}\gamma_n^2(|f(X_{n-1})|^2+|D_n|^2).$$

Let $0 < \delta < 1$ and fix $N \in \mathbb{N}$ such that, for all $n \geq N$, we have $\|f\|_{\mathrm{Lip}(\mathbb{R}^d)}\gamma_n < \delta$. For $n \geq N$, we let

$$\Xi_n = \sum_{\ell=N+1}^{n} \gamma_\ell\big((1 - \|f\|_{\mathrm{Lip}(\mathbb{R}^d)}\gamma_\ell)|f(X_{\ell-1})|^2 + \langle f(X_{\ell-1}), D_\ell\rangle\big)$$

and

$$\Xi'_n = -\|f\|_{\mathrm{Lip}(\mathbb{R}^d)} \sum_{\ell=N+1}^{n} \gamma_\ell^2 |D_\ell|^2$$

and observe that

$$F(X_n) - F(X_N) \geq \Xi_n + \Xi'_n. \tag{D.2}$$

First, suppose that inequality (D.1) is true for all $n > N$. We deduce an estimate for the supremum of the process $(\Xi_n)_{n>N}$. In terms of the martingale

$$(M_n)_{n>N} = \big(-\sum_{\ell=N+1}^{n} \gamma_\ell\langle f(X_{\ell-1}), D_\ell\rangle\big)_{n>N},$$

we have

$$\langle M\rangle_n = \sum_{\ell=N+1}^{n} \gamma_\ell^2\, \mathbb{E}[\langle f(X_{\ell-1}), D_\ell\rangle^2|\mathcal{F}_{\ell-1}] \leq \sum_{\ell=N+1}^{n} C_D^2\gamma_\ell|f(X_{\ell-1})|^2(\sigma_\ell^{\mathrm{RM}})^2.$$

With $\bar\sigma^{\mathrm{RM}} = \sup_{n>N}\sigma_n^{\mathrm{RM}} < \infty$ we get

$$\langle M\rangle_n \leq (C_D\bar\sigma^{\mathrm{RM}})^2 \sum_{\ell=N+1}^{n} \gamma_\ell|f(X_{\ell-1})|^2.$$

Consequently,

$$\Xi_n \geq -\Big(M_n - \frac{1-\delta}{(C_D\bar\sigma^{\mathrm{RM}})^2}\langle M\rangle_n\Big) = -a\Big(\frac{1}{a}M_n - \langle\frac{1}{a}M\rangle_n\Big),$$

for $a := \frac{(C_D\bar\sigma^{\mathrm{RM}})^2}{1-\delta}$. Using Lemma 3.6 in [8], we get that

$$\begin{aligned}\mathbb{P}\Big(\inf_{n>N}\Xi_n \leq -T\Big) &\leq \mathbb{P}\Big(\sup_{n>N}\frac{1}{a}M_n - \langle\frac{1}{a}M\rangle_n \geq \frac{T}{a}\Big)\\ &\leq \frac{4a^2}{T^2} + \sum_{n\in\mathbb{N}_0}\frac{2^{n+3}}{(2^n + \frac{T}{a})^2} \overset{T\to\infty}{\longrightarrow} 0,\end{aligned}$$

so that $\inf_{n>N}\Xi_n$ is almost surely finite. Moreover,

$$\mathbb{E}\Big[\sup_{n>N} -\Xi'_n\Big] = \|f\|_{\mathrm{Lip}(\mathbb{R}^d)} \sum_{\ell=N+1}^{\infty} \gamma_\ell^2\, \mathbb{E}[|D_\ell|^2] \leq \|f\|_{\mathrm{Lip}(\mathbb{R}^d)}C_D^2 \sum_{\ell=N+1}^{\infty} \gamma_\ell(\sigma_\ell^{\mathrm{RM}})^2 < \infty,$$

so that also $\inf_{n>N}\Xi_n$ is finite, almost surely. Using (D.2), we thus get that $\liminf F(X_n) > -\infty$, almost surely, which achieves the proof under the additional assumption that (D.1) is true for all $n > N$.

For the proof of the general result, consider the dynamical system (1.1) with $(D_n)_{n\in\mathbb{N}}$ replaced by $(D_n^{(N)})_{n\in\mathbb{N}}$, given by

$$D_n^{(N)} = \begin{cases} D_n, & \text{if } n \le N \text{ or } \mathbb{E}[|D_n|^2|\mathcal{F}_{n-1}] \le C_D^2(\sigma_n^{\mathrm{RM}})^2/\gamma_n, \\ 0, & \text{else,} \end{cases}$$

for all $n \in \mathbb{N}$. Obviously, the respective N-dependent dynamical system $\mathbf{X}^{(N)} = (X_n^{(N)})_{n\in\mathbb{N}_0}$ satisfies the stronger assumption and we can conclude that

$$\limsup_{n\to\infty} |X_n^{(N)}| < \infty, \quad \text{almost surely.}$$

The result follows since

$$\lim_{N\to\infty} \mathbb{P}(\mathbf{X}^{(N)} \not\equiv \mathbf{X}) = 0$$

as consequence of assumption (D.1). □

References

[1] Arthur E. Albert and Leland A. Gardner, Jr., *Stochastic approximations and nonlinear regression*, MIT Press Research Monographs, No. 42, The M.I.T. Press, Cambridge, Mass., 1967. MR0224217

[2] Albert Benveniste, Michel Métivier, and Pierre Priouret, *Adaptive algorithms and stochastic approximations*, Applications of Mathematics (New York), vol. 22, Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson. MR1082341

[3] Rajendra Bhatia, *Matrix analysis*, Graduate Texts in Mathematics, vol. 169, Springer-Verlag, New York, 1997. MR1477662

[4] Julius R. Blum, *Approximation methods which converge with probability one*, Ann. Math. Statistics **25** (1954), 382–386. MR62399

[5] K. L. Chung, *On a stochastic approximation method*, Ann. Math. Statistics **25** (1954), 463–483. MR64365

[6] Yaim Cooper, *Global minima of overparameterized neural networks*, SIAM J. Math. Data Sci. **3** (2021), no. 2, 676–691. MR4257873

[7] Steffen Dereich, *General multilevel adaptations for stochastic approximation algorithms II: CLTs*, Stochastic Process. Appl. **132** (2021), 226–260. MR4184368

[8] Steffen Dereich and Sebastian Kassing, *Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes*, arXiv:2102.09385, 2021.

[9] Steffen Dereich and Sebastian Kassing, *On minimal representations of shallow ReLU networks*, Neural Networks **148** (2022), 121–128.

[10] C. Derman and J. Sacks, *On Dvoretzky's stochastic approximation theorem*, Ann. Math. Statist. **30** (1959), 601–606. MR107893

[11] Ewa Dudek and Konstanty Holly, *Nonlinear orthogonal projection*, Ann. Polon. Math. **59** (1994), no. 1, 1–31. MR1270298

[12] Marie Duflo, *Algorithmes stochastiques*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 23, Springer-Verlag, Berlin, 1996. MR1612815

[13] Aryeh Dvoretzky, *On stochastic approximation*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I, University of California Press, Berkeley-Los Angeles, Calif., 1956, pp. 39–55. MR0084911

[14] Aryeh Dvoretzky, *Stochastic approximation revisited*, Adv. in Appl. Math. **7** (1986), no. 2, 220–227. MR845378

[15] Paul M. N. Feehan, *Resolution of singularities and geometric proofs of the Łojasiewicz inequalities*, Geom. Topol. **23** (2019), no. 7, 3273–3313. MR4046966

[16] Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen, *Convergence rates for the stochastic gradient descent method for non-convex objective functions*, J. Mach. Learn. Res. **21** (2020), Paper No. 136, 48. MR4138120

[17] Renato Fiorenza, *Hölder and locally Hölder continuous functions, and open sets of class $C^k, C^{k,\lambda}$*, Frontiers in Mathematics, Birkhäuser/Springer, Cham, 2016. MR3588287

[18] Alfred Gray, *Tubes*, second ed., Progress in Mathematics, vol. 221, Birkhäuser Verlag, Basel, 2004. With a preface by Vicente Miquel. MR2024928

[19] P. Hall and C. C. Heyde, *Martingale limit theory and its application*, Probability and Mathematical Statistics, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980. MR624435

[20] Erich Häusler and Harald Luschgy, *Stable convergence and stable limit theorems*, Probability Theory and Stochastic Modelling, vol. 74, Springer, Cham, 2015. MR3362567

[21] Arnulf Jentzen and Philippe von Wurstemberger, *Lower error bounds for the stochastic gradient descent optimization algorithm: sharp convergence rates for slowly and fast decaying learning rates*, J. Complexity **57** (2020), 101438, 16. MR4055054

[22] Harold J. Kushner and G. George Yin, *Stochastic approximation and recursive algorithms and applications*, second ed., Applications of Mathematics (New York), vol. 35, Springer-Verlag, New York, 2003, Stochastic Modelling and Applied Probability. MR1993642

[23] Gunther Leobacher and Alexander Steinicke, *Existence, uniqueness and regularity of the projection onto differentiable manifolds*, Ann. Global Anal. Geom. **60** (2021), no. 3, 559–587. MR4304862

[24] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein, *Visualizing the loss landscape of neural nets*, Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018.

[25] Lennart Ljung, Georg Pflug, and Harro Walk, *Stochastic approximation and optimization of random systems*, DMV Seminar, vol. 17, Birkhäuser Verlag, Basel, 1992. MR1162311

[26] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les Équations aux Dérivées Partielles (Paris, 1962), Éditions du Centre National de la Recherche Scientifique (CNRS), Paris, 1963, pp. 87–89. MR0160856

[27] S. Łojasiewicz, *Ensembles semi-analytiques*, Lectures Notes IHES (Bures-sur-Yvette) (1965).

[28] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher, *On the almost sure convergence of stochastic gradient descent in non-convex problems*, Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 1117–1128.

[29] M. B. Nevel'son and R. Z. Has'minskiĭ, *Stochastic approximation and recursive estimation*, Translations of Mathematical Monographs, Vol. 47, American Mathematical Society, Providence, R.I., 1973. Translated from the Russian by the Israel Program for Scientific Translations. MR0423714

[30] B. T. Polyak, *A new method of stochastic approximation type*, Avtomat. i Telemekh. (1990), no. 7, 98–107. MR1071220

[31] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim. **30** (1992), no. 4, 838–855. MR1167814

[32] Alfréd Rényi, *On stable sequences of events*, Sankhyā Ser. A **25** (1963), 293–302. MR170385

[33] Herbert Robbins and Sutton Monro, *A stochastic approximation method*, Ann. Math. Statistics **22** (1951), 400–407. MR42668

[34] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971), Academic Press, New York, 1971, pp. 233–257. MR0343355

[35] David Ruppert, *Efficient estimators from a slowly convergent robbins-monro procedure*, School of Oper. Res. and Ind. Eng., Cornell Univ., Ithaca, NY, Tech. Rep **781** (1988).

[36] Jerome Sacks, *Asymptotic distribution of stochastic approximation procedures*, Ann. Math. Statist. **29** (1958), 373–405. MR98427

[37] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I. Jordan, *Averaging stochastic gradient descent on Riemannian manifolds*, Proceedings of the 31st Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 75, PMLR, 06–09 July 2018, pp. 650–687.

[38] Rene Vidal, Joan Bruna, Raja Giryes, and Stefano Soatto, *Mathematics of deep learning*, arXiv:1712.04741, 2017.

[39] J. Wolfowitz, *On stochastic approximation methods*, Ann. Math. Statist. **27** (1956), 1151–1156. MR86437

Acknowledgments. The authors would like to thank an anonymous referee for his valuable comments.