Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation

Valentin de Bortoli, Alain Durmus, Marcelo Pereyra, Ana Fernandez Vidal

Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation.

Valentin De Bortoli *[1], Alain Durmus † [1], Marcelo Pereyra ‡ [2], and Ana Fernandez Vidal §[2]

[1]CMLA - École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, 94235 Cachan, France.

[2]School of Mathematical and Computer Sciences, Heriot Watt University & Maxwell Institute for Mathematical Sciences.

May 31, 2020

Abstract

Stochastic approximation methods play a central role in maximum likelihood estimation problems involving intractable likelihood functions, such as marginal likelihoods arising in problems with missing or incomplete data, and in parametric empirical Bayesian estimation. Combined with Markov chain Monte Carlo algorithms, these stochastic optimisation methods have been successfully applied to a wide range of problems in science and industry. However, this strategy scales poorly to large problems because of methodological and theoretical difficulties related to using high-dimensional Markov chain Monte Carlo algorithms within a stochastic approximation scheme. This paper proposes to address these difficulties by using unadjusted Langevin algorithms to construct the stochastic approximation. This leads to a highly efficient stochastic optimisation methodology with favourable convergence properties that can be quantified explicitly and easily checked. The proposed methodology is demonstrated with three experiments, including a challenging application to high-dimensional statistical audio analysis and a sparse Bayesian logistic regression with random effects problem.

1 Introduction

Maximum likelihood estimation (MLE) is central to modern statistical science. It is a cornerstone of frequentist inference [7], and also plays a fundamental role in parametric empirical Bayesian inference [11, 13]. For simple statistical models, MLE can be performed analytically and exactly.

*Email: valentin.debortoli@cmla.ens-cachan.fr

†Email: alain.durmus@cmla.ens-cachan.fr

‡Email: m.pereyra@hw.ac.uk

§Email: af69@hw.ac.uk

However, for most models, it requires using numerical computation methods, particularly optimisation schemes that iteratively seek to maximise the likelihood function and deliver an approximate solution. Following decades of active research in computational statistics and optimisation, there are now several computationally efficient methods to perform MLE in a wide range of classes of models [32, 8].

In this paper we consider MLE in models involving incomplete or "missing" data, such as hidden, latent or unobserved variables, and focus on Expectation Maximisation (EM) optimisation methods [18], which are the predominant strategy in this setting . While the original EM optimisation methodology involved deterministic steps, modern EM methods are mainly stochastic [49]. In particular, they typically rely on a Robbins-Monro stochastic approximation (SA) scheme that uses a Monte Carlo stochastic simulation algorithm to approximate the gradients that drive the optimisation procedure [48, 17, 37, 30]. In many cases, SA methods use Markov chain Monte Carlo (MCMC) algorithms, leading to a powerful general methodology which is simple to implement, has a detailed convergence theory [2], and can address a wide range of moderately low-dimensional models. Alternatively, some stochastic EM schemes use a Gibbs sampling algorithm [12], however this requires running several fully converged MCMC chains and can be significantly more computationally expensive as a result.

The expectations and demands on SA methods constantly rise as we seek to address larger problems and provide stronger theoretical guarantees on the solutions delivered. Unfortunately, existing SA methodology and theory do not scale well to large problems. The reasons are twofold. First, the family of MCMC kernels driving the SA scheme needs to satisfy uniform geometric ergodicity conditions that are usually difficult to verify for high-dimensional MCMC kernels. Second, the existing theory requires using asymptotically exact MCMC methods. In practice, these are usually high-dimensional Metropolis-Hastings methods such as the Metropolis-adjusted Langevin algorithm [51] or Hamiltonian Monte Carlo [33, 22], which are difficult to calibrate within the SA scheme to achieve a prescribed acceptance rate. For these reasons, practitioners rarely use SA schemes in high-dimensional settings.

In this paper, we propose to address these limitations by using inexact MCMC methods to drive the SA scheme, particularly unadjusted Langenvin algorithms, which have easily verifiable geometric ergodicity conditions, and are easy to calibrate [21, 15]. This will allow us to design a high-dimensional stochastic optimisation scheme with favourable convergence properties that can be quantified explicitly and easily checked.

Our contributions are structured as follows: Section 2 formalises the class of MLE problems considered and presents the proposed stochastic optimisation method, which is based on a SA approach driven by an unadjusted Langevin algorithm. Section 3 presents three numerical experiments that demonstrate the proposed methodology in a variety of scenarios. Detailed theoretical convergence results for the method are reported in Section 4, which also describes a generalisation of the proposed methodology and theory to other inexact Markov kernels. The online supplementary material includes additional theoretical results and some details on computational aspects.

2 The stochastic optimisation via unadjusted Langevin method

The proposed Stochastic Optimisation via Unadjusted Langevin (SOUL) method is useful for solving maximum likelihood estimation problems involving intractable likelihood functions. The method is a SA iterative scheme that is driven by an unadjusted Langevin MCMC algorithm. Langevin algorithms are very efficient in high dimensions and lead to an SA scheme that inherits their favourable convergence properties.

2.1 Maximum marginal likelihood estimation

Let Θ be a convex closed set in $\mathbb{R}^{d_\Theta}$. The proposed optimisation method is well-suited for solving maximum likelihood estimation problems of the form

$$\theta^\star \in \underset{\theta\in\Theta}{\arg\max}\ \log p(y|\theta) - g(\theta)\,, \tag{1}$$

where the parameter of interest θ is related to the observed data $y \in \mathsf{Y}$ by a likelihood function $p(y, x|\theta)$ involving an unknown quantity $x \in \mathbb{R}^d$, which is removed from the model by marginalisation. More precisely, we consider problems where the resulting marginal likelihood

$$p(y|\theta) = \int_{\mathbb{R}^d} p(y, x|\theta)\mathrm{d}x\,,$$

is computationally intractable, and focus on models where the dimension of x is large, making the computation of (1) even more difficult. For completeness, we allow the use of a penalty function $g : \Theta \to \mathbb{R}$, or set $g = 0$ to recover the standard maximum likelihood estimator.

As mentioned previously, the maximum marginal likelihood estimation problem (1) arises in problems involving latent or hidden variables [18]. It is also central to parametric empirical Bayes approaches that base their inferences on the pseudo-posterior distribution $p(x|y, \theta^\star) = p(y, x|\theta^\star)/p(y|\theta^\star)$ [11]. Moreover, the same optimisation problem also arises in hierarchical Bayesian maximum-a-posteriori estimation of θ given y, with marginal posterior $p(\theta|y) \propto p(y|\theta)p(\theta)$ where $p(\theta)$ denotes the prior for θ; in that case $g(\theta) = -\log p(\theta)$ [7].

Finally, in this paper we assume that $\log p(y, x|\theta)$ is continuously differentiable with respect to x and θ, and that g is also continuously differentiable with respect to θ. A generalisation of the proposed methodology to non-smooth models is presented in a forthcoming paper [53] that focuses on non-smooth statistical imaging models.

2.2 Stochastic approximation methods

The scheme we propose to solve the optimisation problem (1) is derived in the SA framework [17], which we recall below.

Starting from any $\theta_0 \in \Theta$, SA schemes seek to solve (1) iteratively by computing a sequence $(\theta_n)_{n\in\mathbb{N}}$ associated with the recursion

$$\theta_{n+1} = \Pi_\Theta[\theta_n + \delta_{n+1}(\Delta_{\theta_n} - \nabla g(\theta_n))]\,, \tag{2}$$

where Δ_{θ_n} is some estimator of the intractable gradient $\theta \mapsto \nabla_\theta \log p(y|\theta)$ at θ_n, Π_Θ denotes the projection onto Θ, and $(\delta_n)_{n\in\mathbb{N}^*} \in (\mathbb{R}_+^*)^{\mathbb{N}^*}$ is a sequence of stepsizes. From an optimisation viewpoint, iteration (2) is a stochastic generalisation of the projected gradient ascent iteration [8] for models with intractable gradients. For $n \in \mathbb{N}$, Monte Carlo estimators Δ_{θ_n} for $\nabla_\theta \log p(y|\theta)$ at θ_n are derived from the identity

$$\begin{aligned}\nabla_\theta \log p(y|\theta) &= \int_{\mathbb{R}^d} \frac{\nabla_\theta p(x, y|\theta)}{p(x, y|\theta)} p(x|y, \theta)\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \nabla_\theta \log p(x, y|\theta) p(x|y, \theta)\mathrm{d}x\,,\end{aligned}$$

which suggests to consider

$$\Delta_{\theta_n} = \frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_k^n, y|\theta_n) \;, \tag{3}$$

where $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$ is a sequence of batch sizes and $(X_k^n)_{k\in\{1,\dots,m_n\}}$ is either an exact Monte Carlo sample from $p(x|y,\theta_n) = p(x,y|\theta_n)/p(y|\theta_n)$, or a sample generated by using a Markov Chain targeting this distribution.

Given a sequence $(\theta_n)_{n=1}^N$ generated by using (2), an approximate solution of (1) can then be obtained by calculating, for example, the average of the iterates, i.e.,

$$\hat{\theta}_N = \left\{ \sum_{n=1}^{N} \delta_n \theta_n \right\} \Big/ \left\{ \sum_{n=1}^{N} \delta_n \right\} \;. \tag{4}$$

This estimate converges almost surely to a solution of (1) as $N \to \infty$ provided that some conditions on $p(y|\theta)$, g, $p(x|y,\theta)$, $(\delta_n)_{n\in\mathbb{N}}$, and Δ_{θ_n} are fulfilled. Indeed, following three decades of active research efforts in computational statistics and applied probability, we now have a good understanding of how to construct efficient SA schemes, and the conditions under which these schemes converge (see for example [6, 29, 20, 1, 44, 2]).

SA schemes are successfully applied to maximum marginal likelihood estimation problems where the latent variable x has a low or moderately low dimension. However, they are seldomly used them when x is high-dimensional because this usually requires using high-dimensional MCMC samplers that, unless carefully calibrated, exhibit poor convergence properties. Unfortunately, calibrating the samplers within a SA scheme is challenging because the target density $p(x|y,\theta_n)$ changes at each iteration. As a result, it is, for example, difficult to use Metropolis-Hastings algorithms that need to achieve a prescribed acceptance probability range. Additionally, the conditions for convergence of MCMC SA schemes are often difficult to verify for high-dimensional samplers. For these reasons, practitioners rarely use SA schemes in high-dimensional settings.

As mentioned previously, we propose to address these difficulties by using modern inexact Langevin MCMC samplers to drive (3). These samplers have received a lot of attention in the late because they can exhibit excellent large-scale convergence properties and significantly outperform their Metropolised counterparts (see [23] for an extensive comparison in the context of Bayesian imaging models). Stimulated by developments in high-dimensional statistics and machine learning, we now have detailed theory for these algorithms, including explicit and easily verifiable geometric ergodicity conditions [21, 15, 26, 16]. This will allow us to design a stochastic optimisation scheme with favourable convergence properties that can be quantified explicitly and easily checked.

2.3 Langevin Markov chain Monte Carlo methods

Langevin MCMC schemes to sample from $p(x|y,\theta)$ are based on stochastic continuous dynamics $(\boldsymbol{X}_t^\theta)_{t\geq 0}$ for which the target distribution $p(x|y,\theta)$ is invariant. Two fundamental examples are the Langevin dynamics solution of the following Stochastic Differential Equation (SDE)

$$\mathrm{d}\boldsymbol{X}_t^\theta = -\nabla_x \log p(\boldsymbol{X}_t^\theta|y,\theta)\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t \;, \tag{5}$$

or the kinetic Langevin dynamics solution of

$$\mathrm{d}\boldsymbol{X}_t^\theta = \boldsymbol{V}_t^\theta \;, \qquad \mathrm{d}\boldsymbol{V}_t^\theta = -\nabla_x \log p(\boldsymbol{X}_t^\theta|y,\theta)\mathrm{d}t - \boldsymbol{V}_t^\theta\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t \;,$$

where $(\boldsymbol{B}_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. Under mild assumptions on $p(x|y,\theta)$, these two SDEs admit strong solutions for which $p(x|y,\theta)$ and $p(x,v|y,\theta) = p(x|y,\theta)\exp(-\|v\|^2/2)/(2\pi)^{d/2}$ are the invariant probability measures. In addition, there are detailed explicit convergence results for $(\boldsymbol{X}_t^\theta)_{t\geq 0}$ to $p(x|y,\theta)$, and for $(\boldsymbol{X}_t^\theta, \boldsymbol{V}_t^\theta)_{t\geq 0}$ to $p(x,v|y,\theta)$, under different metrics [25, 24].

However, sampling path solutions for these continuous-time dynamics is not feasible in general. Therefore discretizations have to be used instead. In this paper, we mainly focus on the Euler-Maruyama discrete-time approximation of (5), known as the Unadjusted Langevin Algorithm (ULA) [51], given by

$$X_{k+1} = X_k - \gamma \nabla_x \log p(X_k|y,\theta) + \sqrt{2\gamma} Z_{k+1} \ , \tag{6}$$

where $\gamma > 0$ is the discretization time step and $(Z_k)_{k\in\mathbb{N}^*}$ is a i.i.d. sequence of d-dimensional zero-mean Gaussian random variables with covariance matrix identity. We will use this Markov kernel to drive our SA schemes.

Observe that (6) does not exactly target $p(x|y,\theta)$ because of the bias introduced by the discrete-time approximation. Computational statistical methods have traditionally addressed this issue by complementing (6) with a Metropolis-Hastings correction step to asymptotically remove the bias [51]. This correction usually deteriorates the convergence properties of the chain and may lead to poor non-asymptotic estimation results, particularly in very high-dimensional settings (see for example [23]). However, until recently it was considered that using (6) without a correction step was too risky. Fortunately, recent works have established detailed theoretical guarantees for (6) that do not require using any correction [15, 21]. A main contribution of this work is to extend these guarantees to SA schemes that are driven by these highly efficient but inexact samplers.

2.4 The SOUL algorithm

We are now ready to present the proposed Stochastic Optimization via Unadjusted Langevin (SOUL) methodology. Let $(\delta_n)_{n\in\mathbb{N}^*} \in (\mathbb{R}_+^*)^{\mathbb{N}^*}$ and $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$ be the sequences of step-sizes and batch sizes defining the SA scheme (2)-(3). For any $\theta\in\Theta$ and $\gamma>0$, denote by $\mathrm{R}_{\gamma,\theta}$ the Langevin Markov kernel (6) to approximately sample from $p(x|y,\theta)$, and by $(\gamma_n)_{n\in\mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ be the sequence of discrete time steps used.

Formally, starting from some $X_0^0 \in \mathbb{R}^d$ and $\theta_0 \in \Theta$, for $n\in\mathbb{N}$ and $k\in\{0,\ldots,m_n-1\}$, we recursively define $(\{X_k^n : k\in\{0,\ldots,m_n\}\},\theta_n)_{n\in\mathbb{N}}$ on a probability space $(\Omega,\mathcal{F},\mathbb{P})$, where $(X_k^n)_{k\in\{0,\ldots,m_n\}}$ is a Markov chain with Markov kernel $\mathrm{R}_{\gamma_n,\theta_n}$, $X_0^n = X_{m_{n-1}}^{n-1}$ given $\mathcal{F}_{n-1}$, and

$$\theta_{n+1} = \Pi_\Theta\left[\theta_n - \frac{\delta_{n+1}}{m_n}\sum_{k=1}^{m_n}\Delta_{\theta_n}(X_k^n)\right] \ ,$$

where we recall that Π_Θ is the projection onto Θ, and for all $n\in\mathbb{N}$

$$\mathcal{F}_n = \sigma\left(\theta_0, \{(X_k^\ell)_{k\in\{0,\ldots,m_\ell\}} : \ell\in\{0,\ldots,n\}\}\right) \ , \qquad \mathcal{F}_{-1} = \sigma(\theta_0) \tag{7}$$

Note that such a construction is always possible by Kolmogorov extension theorem [34, Theorem 5.16], hence for any $n\in\mathbb{N}$, θ_{n+1} is $\mathcal{F}_n$-measurable. Then, as mentioned previously, we compute a sequence of approximate solutions of (1) by calculating, for example,

$$\hat{\theta}_N = \left\{\sum_{n=1}^N \delta_n\theta_n\right\} \Bigg/ \left\{\sum_{n=1}^N \delta_n\right\} \ . \tag{8}$$

The pseudocode associated with the proposed SOUL method is presented in Algorithm 1 below. Observe that, for additional efficiency, instead of generating independent Markov chains at each SA iteration, we warm-start the chains by setting $X_0^n = X_{m_{n-1}}^{n-1}$, for any $n \in \{1, \ldots, N\}$.

Algorithm 1 The Stochastic Optimization via Unadjusted Langevin (SOUL) method

Inputs:
  $\theta_0 \in \Theta$, $X_0^0 \in \mathbb{R}^d$, $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$, $(m_n)_{n\in\mathbb{N}}$, N
for $n \in \{1, \ldots, N-1\}$ **do**
  if $n \geq 1$ **then**
    $X_0^n = X_{m_{n-1}}^{n-1}$
  end if
  for $k \in \{0, \ldots, m_n - 1\}$ **do**
    $Z_{k+1}^n \sim \mathrm{N}(0, \mathrm{I}_d)$
    $X_{k+1}^n = X_k^n + \gamma_n \nabla_x \log p(X_k^n | y, \theta_n) + \sqrt{2\gamma_n} Z_{k+1}^n$
  end for
  $\Delta_{\theta_n} = \frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_k^n, y | \theta_n)$
  $\theta_{n+1} = \Pi_\Theta[\theta_n + \delta_{n+1}(\Delta_{\theta_n} - \nabla g(\theta_n))]$
end for
Outputs:
  $\hat{\theta}_N = \left\{\sum_{n=1}^N \delta_n \theta_n\right\} \Big/ \left\{\sum_{n=1}^N \delta_n\right\}$

To conclude, Section 3 below demonstrates the proposed methodology with three numerical experiments related to high-dimensional logistic regression and statistical audio analysis with sparsity promoting priors. A detailed theoretical analysis of the proposed SOUL method is reported in Section 4. More precisely, we establish that if the cost function $f(\theta) = g(\theta) - \log p(y|\theta)$ defining (1) is convex, and if $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ go to 0 sufficiently fast, then $\mathbb{E}[f(\hat{\theta}_N)]$ converges to $\min_\Theta f$ and quantify the rate of convergence. Moreover, in the case where $(\gamma_n)_{n\in\mathbb{N}}$ is held fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma$, we show convergence to a neighbourhood of the solution, in the sense that there exist explicit $C, \alpha > 0$ such that $\limsup_{N\to+\infty} \mathbb{E}[f(\hat{\theta}_N)] - \min_\Theta f \leq C\gamma^\alpha$. Finally, we also study the important case where f is not convex. In that case, we use the results of [37] to establish that $(\theta_n)_{n\in\mathbb{N}}$ converges almost surely to a stationary point of the projected ordinary differential equation associated with ∇f and Θ. We postpone this result to Appendix B in the supplementary document because it is highly technical.

3 Numerical results

We now demonstrate the proposed methodology with three experiments that we have chosen to illustrate a variety of scenarios. Section 3.1 presents an application to empirical Bayesian logistic regression, where (1) can be analytically shown to be a convex optimisation problem with an unique solution $\theta^\star$, and where we benchmark our MLE estimate against the solution obtained by calculating the marginal likelihood $p(y|\theta)$ over a θ-grid by using an harmonic mean estimator. Furthermore, Section 3.2 presents a challenging application related to statistical audio compressive sensing analysis, where we use SOUL to estimate a regularisation parameter that controls the degree of sparsity enforced, and where a main difficulty is the high-dimensionality of the latent space ($d = 2,900$). Finally, Section 3.3 presents an application to a high-dimensional empirical

Bayesian logistic regression with random effects for which the optimisation problem (1) is not convex. All experiments were carried out on an Intel i9-8950HK@2.90GHz workstation running Matlab R2018a.

3.1 Bayesian Logistic Regression

In this first experiment we illustrate the proposed methodology with an empirical Bayesian logistic regression problem [55, 45]. We observe a set of covariates $\{v_i\}_{i=1}^{d_y} \in \mathbb{R}^d$, and binary responses $\{y_i\}_{i=1}^{d_y} \in \{0,1\}$, which we assume to be conditionally independent realisations of a logistic regression model: for any $i \in \{1,\dots,d_y\}$, y_i given β and v_i has distribution $\mathrm{Ber}(s(v_i^{\mathrm{T}}\beta))$, where $\beta \in \mathbb{R}^d$ is the regression coefficient, $\mathrm{Ber}(\alpha)$ denotes the Bernoulli distribution with parameter $\alpha \in [0,1]$ and $s(u) = \mathrm{e}^u/(1+\mathrm{e}^u)$ is the cumulative distribution function of the standard logistic distribution. The prior for β is set to be $\mathrm{N}(\theta\mathbf{1}_d, \sigma^2\,\mathrm{I}_d)$, the d-dimensional Gaussian distribution with mean $\theta\mathbf{1}_d$ and covariance matrix $\sigma^2\,\mathrm{I}_d$, where θ is the parameter we seek to estimate, $\mathbf{1}_d = (1,\dots,1) \in \mathbb{R}^d$, $\sigma^2 = 5$ and I_d is the d-dimensional identity matrix. Following an empirical Bayesian approach, the parameter θ is computed by maximum marginal likelihood estimation using Algorithm 1 with the marginal likelihood $p(y|\theta)$ given by

$$p(y|\theta) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \left\{ \prod_{i=1}^{d_y} s(v_i^{\mathrm{T}}\beta)^{y_i} (1 - s(v_i^{\mathrm{T}}\beta))^{1-y_i} \right\} \mathrm{e}^{-\frac{\|\beta-\theta\mathbf{1}_d\|^2}{2\sigma^2}} \mathrm{d}\beta \,. \tag{9}$$

Lemma 7 in Appendix A of the supplementary document shows that (9) is log-concave with respect to θ. We use the proposed SOUL methodology to estimate $\theta^\star$ for the Wisconsin Diagnostic Breast Cancer dataset[1], for which $d_y = 683$ and $d = 10$, and where we suitably normalise the covariates. In order to assess the quality of our estimation results, we also calculate $p(y|\theta)$ over a grid of values for θ by using a truncated harmonic mean estimator.

To implement Algorithm 1 we derive the log-likelihood function

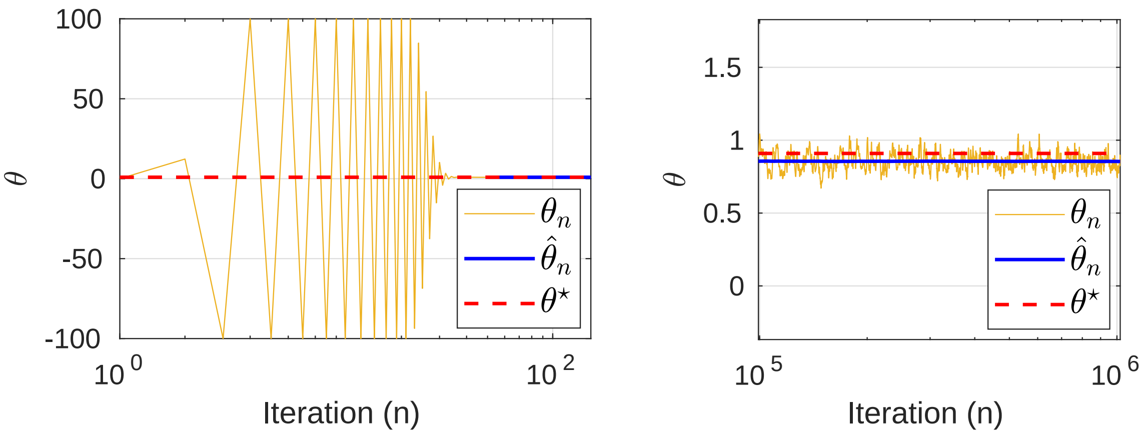
$$\log p(y|\beta,\theta) = \sum_{i=1}^{d_y} \left\{ y_i v_i^{\mathrm{T}}\beta - \log(1 + \mathrm{e}^{(v_i^{\mathrm{T}}\beta)}) \right\} \,,$$

and obtain the following expressions for the gradients used in the MCMC steps (6) and SA steps (2) respectively

$$\nabla_\beta \log p(\beta|y,\theta) = \sum_{i=1}^{d_y} \left\{ y_i v_i - s(v_i^{\mathrm{T}}\beta) v_i \right\} - \frac{(\beta - \theta\mathbf{1}_d)}{\sigma^2} \,,$$
$$\nabla_\theta p(\beta, y|\theta) = \langle \mathbf{1}_d, \beta - \theta\mathbf{1}_d \rangle / \sigma^2 \,.$$

For the MCMC steps, we use a fixed stepsize $\gamma_n = 8.34 \times 10^{-5}$, and batch size $m_n = 1$, for any $n \in \mathbb{N}$. On the other hand, we consider for the SA steps, the sequence of stepsizes $\delta_n = 60/n^{0.8}$, $\Theta = [-100, 100]$ and $\theta_0 = 0$. Finally, we first run 100 burn-in iterations with fixed $\theta_n = \theta_0$ to warm-up the Markov chain, followed by 50 iterations of Algorithm 1 to warm-up the iterates. This procedure is then followed by $N = 10^6$ iterations of Algorithm 1 to compute $\hat{\theta}_N$.

[1] Available online: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)



(a) (b)

Figure 1: Bayesian logistic regression - Evolution of the iterates $\hat{\theta}_n$ and θ_n for the proposed method during (a) burn-in phase and (b) convergence phase. An estimate of $\theta^\star$, the true maximiser of $p(y|\theta)$, is plotted as a reference.

Figure 1(a) shows the evolution of the iterates θ_n during the first 100 iterations. Observe that the sequence initially oscillates, and then stabilises close to $\theta^\star$ after approximately 50 iterations. Figure 1(b) presents the iterates θ_n for $n = 10^5, \ldots, 10^6$. For completeness, Figure 2 shows the histograms corresponding to the marginal posteriors $p(\beta_j|y, v, \hat{\theta}_N)$, for $j = 1, \ldots, 10$, obtained as a by-product of Algorithm 1. In order to verify that the obtained estimate $\hat{\theta}_N$ is close to the

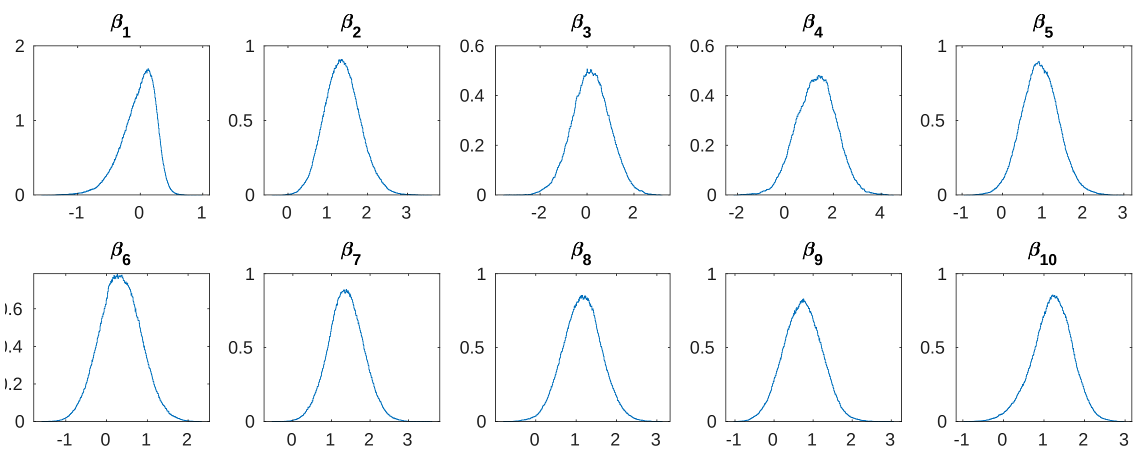


Figure 2: Bayesian logistic regression - Normalised histograms of each component of β obtained with 2×10^6 Monte Carlo samples.

true MLE $\theta^\star$ we use a truncated harmonic mean estimator (THME) [50] to calculate the marginal likelihood $p(y|\theta)$ for a range of values of θ. Although obtaining the THME is usually computationally expensive, it is viable in this particular experiment as β is low-dimensional. More precisely, given n samples $(\beta_i)_{i\in\{1,\dots,n\}}$ from $p(\beta|y,\theta)$, we obtain an approximation of $p(y|\theta)$ by computing

$$\hat{p}(y|\theta) = n\,\mathrm{Vol}(\mathsf{A}) \Bigg/ \left(\sum_{k=1}^{n} \frac{\mathbb{1}_{\mathsf{A}}(\beta_k)}{p(\beta_k, y|\theta)}\right) \;,$$

where A is a d-dimensional ball centered at the posterior mean $\bar{\beta} = n^{-1}\sum_{k=1}^{n}\beta_k$, and with radius set such that $n^{-1}\sum_{i=1}^{n}\mathbb{1}_{\mathsf{A}}(\beta_i) \approx 0.4$. Using $n = 6\times 10^5$ samples, we obtain the approximation shown in Figure 3(a), where in addition to the estimated points we also display a quadratic fit (corresponding to a Gaussian fit in linear scale), which we use to obtain an estimate of $\theta^\star$ (the obtained log-likelihood values are small because the dataset is large ($d_y = 683$)).

To empirically study the estimation error involved, we replicate the experiment 10^3 times. Figure 3 shows the obtained histogram of $\{\hat{\theta}_{N,i}\}_{i=1}^{1000}$, where we observe that all these estimators are very close to the true maximiser $\theta^\star$. Besides, note that the distribution of the estimation error is close to a Gaussian distribution, as expected for a maximum likelihood estimator. Also, there is a small estimation bias of the order of 3%, which can be attributed to the discretization error of SDE (5), and potentially to a small error in the estimation of $\theta^\star$.

We conclude this experiment by using SOUL to perform a predictive empirical Bayesian analysis on the binary responses. We split the original dataset into an 80% training set $(y^{\mathrm{train}}, v^{\mathrm{train}})$ of size $d_{\mathrm{train}} = 546$, and a 20% test set $(y^{\mathrm{test}}, v^{\mathrm{test}})$ of size $d_{\mathrm{test}} = 137$, and use SOUL to draw samples from the predictive distribution $p(y^{\mathrm{test}}|y^{\mathrm{train}}, v^{\mathrm{train}}, v^{\mathrm{test}}, \hat{\theta}_N)$. More precisely, we use SOUL to simultaneously calculate $\hat{\theta}_N$ and simulate from $p(\beta|y^{\mathrm{train}}, v^{\mathrm{train}}, \hat{\theta}_N)$, followed by simulation from $p(y^{\mathrm{test}}|\beta, y^{\mathrm{train}}, v^{\mathrm{train}}, v^{\mathrm{test}})$. We then estimate the maximum-a-posteriori predictive response $\hat{y}^{\mathrm{test}}$, and measure prediction accuracy against the test dataset by computing the error

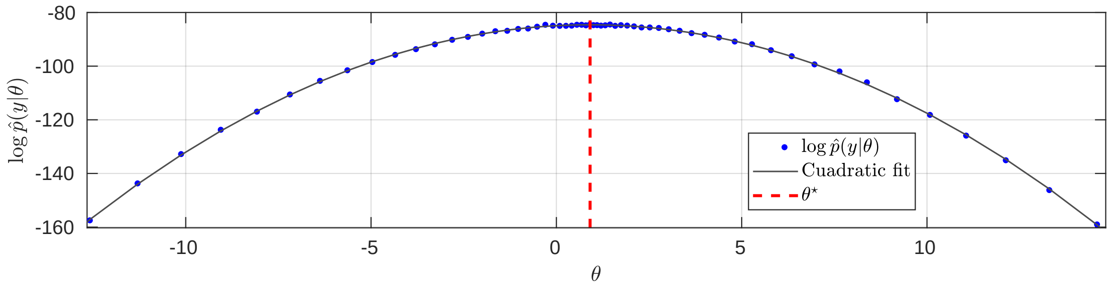
$$\epsilon = \|y^{\mathrm{test}} - \hat{y}^{\mathrm{test}}\|_1 / d_{\mathrm{test}} = \sum_{i=1}^{d_{\mathrm{test}}} \left|y_i^{\mathrm{test}} - \hat{y}_i^{\mathrm{test}}\right| / d_{\mathrm{test}} \;,$$

and obtain $\epsilon = 2.2\%$. For comparison, Figure 4 below reports the error ϵ as a function of θ (the discontinuities arise because of the highly non-linear nature of the model). Observe that the estimated $\hat{\theta}_N$ produces a model that has a very good performance in this regard.

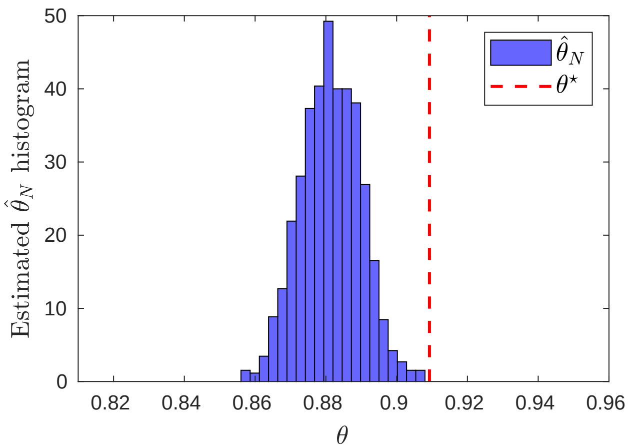
3.2 Statistical audio compression

Compressive sensing techniques exploit sparsity properties in the data to estimate signals from fewer samples than required by the Nyquist–Shannon sampling theorem [10, 9]. Many real-world data admit a sparse representation on some basis or dictionary. Formally, consider an ℓ-dimensional time-discrete signal $z \in \mathbb{R}^\ell$ that is sparse in some dictionary $\mathbf{\Psi} \in \mathbb{R}^{\ell\times d}$, i.e, there exists a latent vector $x \in \mathbb{R}^d$ such that $z = \mathbf{\Psi}x$ and $\|x\|_0 = \sum_{i=1}^{d}\mathbb{1}_{\mathbb{R}^*}(x_i) \ll \ell$. This prior assumption can be modelled by using a smoothed-Laplace distribution [38]

$$p(x|\theta) \propto \exp\left(-\theta\sum_{i=1}^{d} h_\lambda(x_i)\right) \;, \tag{10}$$



(a)



(b)

Figure 3: Bayesian logistic regression - (a) Estimated points of the marginal log-likelihood $\log \hat{p}(y|\theta)$ with quadratic fit (corresponding to a Gaussian fit in linear scale). (b) Normalised histogram of $\hat{\theta}_N$ for 1000 repetitions of the experiment. An estimate of $\theta^\star$, the maximiser of $\hat{p}(y|\theta)$, is plotted as a reference.

where h_λ is the Huber function given for any $u \in \mathbb{R}$ by

$$h_\lambda(u) = \begin{cases} u^2/2 & \text{if } |u| \leq \lambda \ , \\ \lambda(|u| - \lambda/2) & \text{otherwise} \ . \end{cases} \tag{11}$$

Acquiring z directly would call for measuring ℓ univariate components. Instead, a carefully designed measurement matrix $\mathbf{M} \in \mathbb{R}^{p \times \ell}$, with $p \ll \ell$, is used to directly observe a "compressed" signal $\mathbf{M}z$, which only requires taking p measurements. In addition, measurements are typically noisy which results in an observation $y \in \mathbb{R}^p$ modeled as $y = \mathbf{M}z + w$ where we assume that the noise w has distribution $\mathrm{N}(0, \sigma^2 \, \mathrm{I}_p)$, and therefore the likelihood function is given by

$$p(y|x) \propto \exp\left(- \|y - \mathbf{M}\boldsymbol{\Psi}x\|_2^2 / (2\sigma^2)\right) \ ,$$

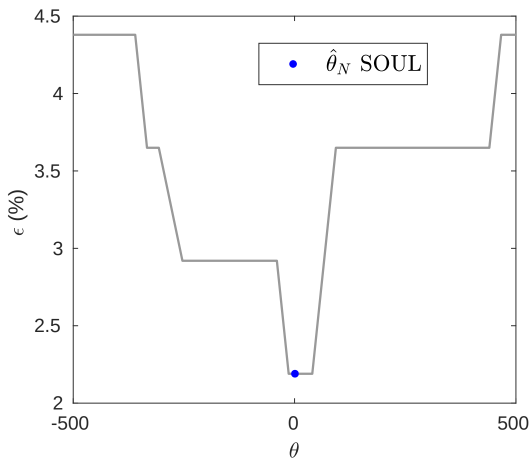


Figure 4: Bayesian logistic regression - Percentage of mislabelled binary observations in terms of θ. In blue we show the value of $\hat{\theta}_N$ obtained with Algo. 1.

leading to the posterior distribution

$$p(x|y) \propto \exp\left(-\|y - \mathbf{M}\boldsymbol{\Psi}x\|_2^2 / (2\sigma^2) - \theta \sum_{i=1}^{d} h_\lambda(x_i)\right) .$$

To recover z from y, we then compute the maximum-a-posteriori estimate

$$\hat{x}_{\mathrm{MAP}} \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \|y - \mathbf{M}\boldsymbol{\Psi}x\|_2^2 / 2\sigma^2 + \theta \sum_{i=1}^{d} h_\lambda(x_i) \right\} , \tag{12}$$

and set $\hat{z}_{\mathrm{MAP}} = \boldsymbol{\Psi}\hat{x}_{\mathrm{MAP}}$.

Following decades of active research, there are now many convex optimisation algorithms that can be used to efficiently solve (12), even when d is very large [14, 43]. However, the selection of the value of θ in (12) remains a difficult open problem. This parameter controls the degree of sparsity of x and has a strong impact on estimation performance.

A common heuristic within the compressive sensing community is to set $\theta_{\mathrm{cs}} = 0.1 \times \|(\mathbf{M}\boldsymbol{\Psi})^\intercal y\|_\infty / \sigma^2$, where for any $z \in \mathbb{R}^\ell$, $\|z\|_\infty = \max_{i \in \{1,\ldots,\ell\}} |z_i|$, as suggested in [35] and [28]; however, better results can arguably be obtained by adopting a statistical approach to estimate θ.

The Bayesian framework offers several strategies for estimating θ from the observation y. In this experiment we adopt an empirical Bayesian approach and use SOUL to compute the MLE $\theta^\star$, which is challenging given the high-dimensionality of the latent space.

To illustrate this approach, we consider the audio experiment proposed in [5] for the *"Mary had a little lamb"* song. The MIDI-generated audio file z has $\ell = 319,725$ samples, but we only have access to a noisy observation vector y with $p = 456$ random time points of the audio signal, corrupted by additive white Gaussian noise with $\sigma = 0.015$. The latent signal x has dimension $d = 2,900$ and is related to z by a dictionary matrix $\boldsymbol{\Psi}$ whose row vectors correspond to different piano notes lasting a quarter-second long [2]. The parameter λ for the prior (10) is set to $\lambda = 4 \times 10^{-5}$. We used the heuristic θ_{cs} as the initial value for θ in our algorithm. To solve the optimisation problem (12) we use the Gradient Projection for Sparse Reconstruction (GPSR) algorithm proposed in [28]. We use this solver because it is the one used in the online MATLAB demonstration of [5], however, more modern algorithms could be used as well. We implemented Algorithm 1 using a fixed stepsize $\gamma_n = 6.9 \times 10^{-6}$, a fixed batch size $m_n = 1$, $\delta_n = 20\, n^{-0.8}/d = 0.0069\, n^{-0.8}$ and 100 burn-in iterations.

The algorithm converged in approximately 500 iterations, which were computed in only 325 milliseconds. Figure 5 (left), shows the first 250 iterations of the sequence θ_n and of the weighted average $\hat{\theta}_n$. Again, observe that the iterates oscillate for a few iterations and then quickly stabilise. Finally, to assess the quality of the estimate $\hat{\theta}_N$, Figure 5 (right) presents the reconstruction mean squared error as a function of θ. The error is measured with respect to the reconstructed signal and is given by $\text{MSE}(\hat{x}_{\text{MAP}}) = \|z^\star - \mathbf{\Psi}\hat{x}_{\text{MAP}}\|_2^2/\ell$, where $z^\star$ is the true audio signal. Observe that the estimated value $\hat{\theta}_N$ is very close to the value that minimises the estimation error, and significantly outperforms the heuristic value θ_{cs} commonly used by practitioners.

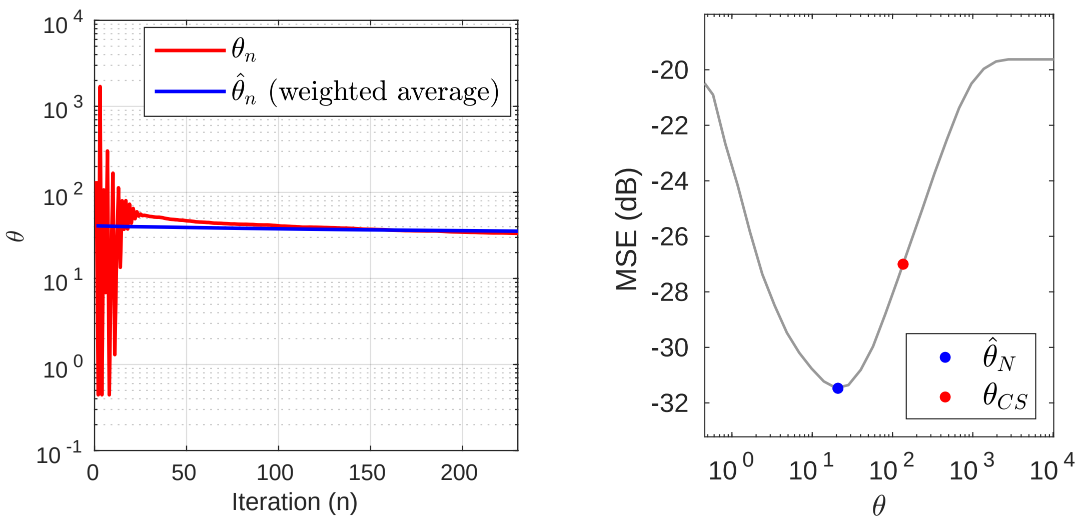


Figure 5: Statistical audio compression - Evolution of the the iterate θ_n and $\hat{\theta}_n$ with $\sigma = 0.015$ in log scale (left). Reconstruction mean squared error (MSE) in dB as a function of the θ (right).

3.3 Sparse Bayesian logistic regression with random effects

Following on from the Bayesian logistic regression in Section 3.1, where $p(y|\theta)$ is log-concave and hence $\theta^\star$ unique, we now consider a significantly more challenging sparse Bayesian logistic regression with random effects problem. In this experiment $p(y|\theta)$ is no longer log-concave, so SOUL can potentially get trapped in local maximisers. Furthermore, the dimension of θ in this experiment

[2] Each quarter-second sound can have one of 100 possible frequencies and be in 29 different positions in time.

is very large ($d_\theta = 1001$), making the MLE problem even more challenging. This experiment was previously considered by [2] and we replicate their setup.

Let $\{y_i\}_{i=1}^{d_y} \in \{0, 1\}$ be a vector of binary responses which can be modelled as d_y conditionally independent realisations of a random effect logistic regression model,

$$y_i|x \sim \mathrm{Ber}\left(s(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x)\right) \ , \quad i \in \{1, \ldots, d_y\} \ ,$$

where $v_i \in \mathbb{R}^p$ are the covariates, $\beta \in \mathbb{R}^p$ is the regression vector, $z_i \in \mathbb{R}^d$ are (known) loading vectors, x are random effects and $\sigma > 0$. In addition, recall that $\mathrm{Ber}(\alpha)$ denotes the Bernoulli distribution with parameter $\alpha \in [0, 1]$ and $s(u) = \mathrm{e}^u/(1+\mathrm{e}^u)$ is the cumulative distribution function of the standard logistic distribution. The goal is to estimate the unknown parameters $\theta = (\beta, \sigma) \in \mathbb{R}^p \times (0, +\infty)$ directly from $\{y_i\}_{i=1}^{d_y}$, without knowing the value of x, which we assume to follow a standard Gaussian distribution, *i.e.* $p(x) = \exp\{-\|x\|_2^2/2\}/(2\pi)^{d/2}$. We estimate θ by MLE using Algorithm 1 to maximize (1), with marginal likelihood given by

$$p(y|\theta) = \int_{\mathbb{R}^d} \prod_{i=1}^{d_y} s(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x)^{y_i}(1 - s(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x))^{1-y_i} p(x)\mathrm{d}x \ ,$$

and we use the penalty function

$$g(\theta) = \sum_{j=1}^{d} h_\lambda(\beta_j) \ , \tag{13}$$

where h_λ is the Huber function defined in (11).

We follow the procedure described in [2] to generate the observations $\{y_i\}_{i=1}^{d_y}$, with $d_y = 500$, $p = 1000$ and $d = 5$[3]. The vector of regressors β_{true} is generated from the uniform distribution on $[1, 5]$ and 98% of its coefficients are randomly set to zero. The variance σ_{true} of the random effect is set to 0.1, and the projection interval for the estimated σ is $[10^{-5}, +\infty)$. Finally, the parameter λ in (13) is set to $\lambda = 30$. We emphasize at this point that θ is high-dimensional in this experiment ($d_\Theta = 1001$), making the estimation problem particularly challenging.

The conditional log-likelihood function for this model is

$$\log p(y|x, \theta) = \sum_{i=1}^{d_y} \left\{ y_i(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x) - \log(1 + \mathrm{e}^{v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x}) \right\} \ .$$

To implement Algorithm 1 we use the gradients

$$\nabla_x \log p(x|y, \theta) = \sum_{i=1}^{d_y} \left\{ \sigma z_i(y_i - s(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x)) \right\} - x \ ,$$

$$\nabla_\theta \log p(x, y|\theta) = \sum_{i=1}^{d_y} \left\{ (y_i - s(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x)) \begin{bmatrix} v_i \\ z_i^{\mathrm{T}}x \end{bmatrix} \right\} \ .$$

Finally the gradient of the penalty function is given by

$$\frac{\partial}{\partial \beta_i} g(\theta) = \begin{cases} \beta_i & |\beta_i| \le \lambda \\ \lambda \ \mathrm{sign}(\beta_i), & |\beta_i| > \lambda \end{cases} , \qquad \frac{\partial}{\partial \sigma} g(\theta) = 0 \ ,$$

[3]We renamed some symbols for notation consistency. What we denote by v_i, x, d_y and d, is denoted in [2] by x_i, $\mathbf{U}$, N and q respectively.

where sign denotes the sign function, *i.e.* for any $s \in \mathbb{R}$, $\text{sign}(s) = |s|/s$ if $s \neq 0$, and $\text{sign}(s) = 0$ otherwise.

We use $\gamma_n = 0.01$, $\delta_n = n^{-0.95}/d = 0.2 \times n^{-0.95}$, a fixed batch size $m_n = 1$, $\beta_0 = \mathbf{1}_p$ and $\sigma_0 = 1$ as initial values. Moreover, we perform 10^4 burn-in iterations with a fixed value of $\theta_0 = (\beta_0, \sigma_0)$ to warm-up the Markov chain, and further 600 iterations of Algorithm 1 to warm-start the iterates. Following on from this, we run $N = 5 \times 10^4$ iterations of Algorithm 1 to compute $\hat{\theta}_N$. Computing this estimates required 25 seconds in total.

Figure 6 shows the evolution of the iterates throughout iterations, where we used $\|\hat{\beta}_n\|_0$ as a summary statistic to track the number of active components. Because the Huber penalty (11) does not enforce exact sparsity on β, to estimate the number of active components we only consider values that are larger than a threshold τ (we used $\tau = 0.005$).

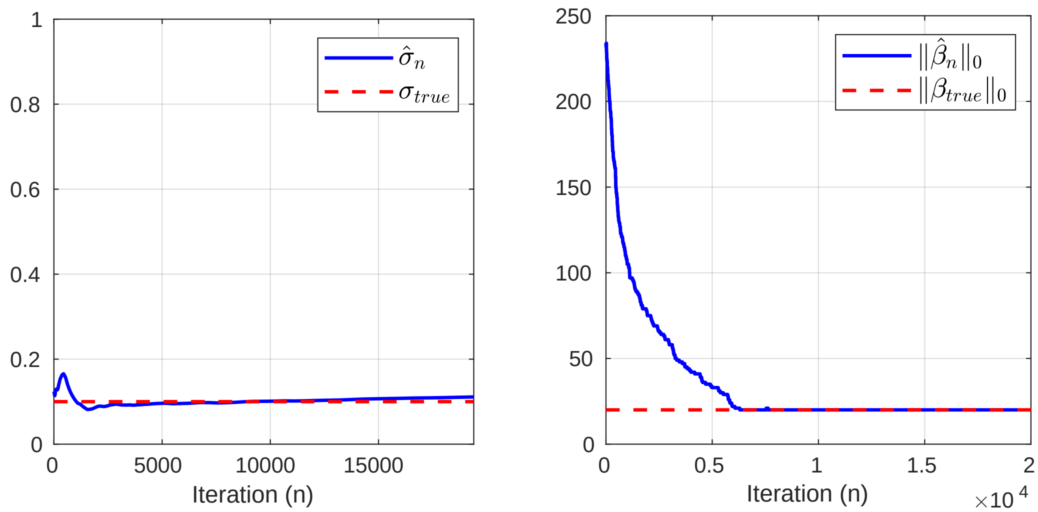


Figure 6: Sparse Bayesian logistic regression with random effects - Evolution of the $\|\hat{\beta}_n\|_0$ and of the iterate $\hat{\sigma}_n$ for the proposed method. The true values are plotted in red as a reference.

From Figure 6 we observe that $\hat{\sigma}_n$ converges to a value that is very close to σ_{true}, and that the number of active components is also accurately estimated. Moreover, Figure 7 shows that most active components were correctly identified. We also observe that $\hat{\beta}_n$ stabilizes after approximately 6300 iterations, which correspond to 6300 Monte Carlo samples as m_n=1. This is in close agreement with the results presented in [2, Figure 5], where they observe stabilization after a similar number of iterations of their highly specialised Polya-Gamma sampler.

It is worth emphasising at this point that [2] considers the non-smooth penalty $g(\theta) = \lambda\|\beta\|_1$ instead of (13). Consequently, instead of using the gradient of g, they resort to the so-called proximal operator of g [14]. The generalisation of the SOUL methodology proposed in this paper to models that have non-differentiable terms is addressed in Vidal and Pereyra [54], Vidal et al. [53].

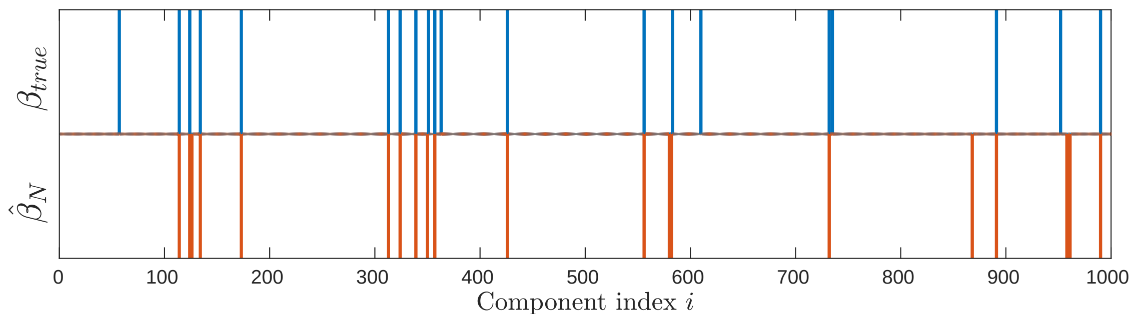


Figure 7: Sparse Bayesian logistic regression with random effects - Support of the estimated $\hat{\beta}_N$ compared with the support of β_{true}.

4 Theoretical convergence analysis for SOUL, and generalisation to other inexact MCMC kernels (SOUK)

In this section we state our main theoretical results for SOUL. For completeness, we first present the results in a general stochastic optimisation setting and by considering a generic inexact MCMC sampler, and then show that our results apply to the specific MLE optimisation problem (1), and to the specific Langevin algorithm (6) used in SOUL.

4.1 Notations and convention

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$, $\mathbb{F}(\mathbb{R}^d)$ the set of all Borel measurable functions on $\mathbb{R}^d$ and for $f \in \mathbb{F}(\mathbb{R}^d)$, $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{F}(\mathbb{R}^d)$ a μ-integrable function, denote by $\mu(f)$ the integral of f with respect to μ. For $f \in \mathbb{F}(\mathbb{R}^d)$, the V-norm of f is given by $\|f\|_V = \sup_{x\in\mathbb{R}^d} |f(x)|/V(x)$. Let ξ be a finite signed measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The V-total variation distance of ξ is defined as

$$\|\xi\|_V = \sup_{f\in\mathbb{F}(\mathbb{R}^d), \|f\|_V \le 1} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}\xi(x) \right| .$$

If $V \equiv 1$, then $\|\cdot\|_V$ is the total variation denoted by $\|\cdot\|_{\mathrm{TV}}$. Let μ be a finite signed measure, then by the Hahn-Jordan theorem [19, Theorem D.1.3], there exists a pair of finite singular measures μ^+, μ^- such that $\mu = \mu^+ - \mu^-$. The total variation measure $|\mu|$ is given by $|\mu| = \mu^+ + \mu^-$.

Let U be an open set of $\mathbb{R}^d$. We denote by $\mathrm{C}^k(\mathsf{U}, \mathbb{R}^p)$ the set of $\mathbb{R}^p$-valued k-differentiable functions, respectively the set of compactly supported $\mathbb{R}^p$-valued k-differentiable functions. $\mathrm{C}^k(\mathsf{U})$ stands $\mathrm{C}^k(\mathsf{U}, \mathbb{R})$. Let $f : \mathsf{U} \to \mathbb{R}$, we denote by ∇f, the gradient of f if it exists. f is said to me m-convex with $m \ge 0$ if for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$,

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - (m/2)\|x - y\|^2 .$$

We recall that if $f : \mathsf{U} \to \mathbb{R}$ is twice differentiable at point $a \in \mathbb{R}^d$, its Laplacian is given by $\Delta f(a) = \sum_{i=1}^d (\partial^2 f)/(\partial x_i^2)(a)$. For any $\mathsf{A} \subset \mathbb{R}^d$, we denote by $\partial\mathsf{A}$ the boundary of A. Let $(\Omega, \mathcal{F}, \mathbb{P})$

be a probability space. Denote by $\mu \ll \nu$ if μ is absolutely continuous with respect to ν and $\mathrm{d}\mu/\mathrm{d}\nu$ an associated density. Let μ, ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the Kullback-Leibler divergence of μ from ν by

$$\mathrm{KL}\left(\mu|\nu\right) = \begin{cases} \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu(x) \;, & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise .} \end{cases}$$

The complement of a set $\mathsf{A} \subset \mathbb{R}^d$, is denoted by A^c. We take the convention that $\prod_{k=p}^n = 1$ and $\sum_{k=p}^n = 0$ for $n, p \in \mathbb{N}$, $n < p$. All densities are w.r.t. the Lebesgue measure unless stated otherwise.

4.2 Stochastic Optimization with inexact MCMC methods

We consider the problem of minimizing a function $f : \Theta \to \mathbb{R}$ with $\Theta \subset \mathbb{R}^{d_\Theta}$ under the following assumptions.

A 1. *Θ is a convex compact set and $\Theta \subset \overline{\mathrm{B}}(0, M_\Theta)$ with $M_\Theta > 0$.*

A 2. *There exist an open set $\mathsf{U} \subset \mathbb{R}^{d_\Theta}$ and $L_f \geq 0$ such that $\Theta \subset \mathsf{U}$, $f \in \mathrm{C}^1(\mathsf{U}, \mathbb{R})$ and satisfies for any $\theta_1, \theta_2 \in \Theta$*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L_f \|\theta_1 - \theta_2\| \;.$$

A 3. *For any $\theta \in \Theta$, there exist $H_\theta :\ \mathbb{R}^d \to \mathbb{R}^{d_\Theta}$ and a probability distribution π_θ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ satisfying that $\pi_\theta(H_\theta) < +\infty$ and for any $\theta \in \Theta$*

$$\nabla f(\theta) = \int_{\mathbb{R}^d} H_\theta(x) \mathrm{d}\pi_\theta(x) \;.$$

In addition, $(\theta, x) \mapsto H_\theta(x)$ is measurable.

Note that for the maximum marginal likelihood estimation problem (1), f corresponds to $\theta \mapsto -\log(p(y|\theta)) + g(\theta)$, for any $\theta \in \Theta$, $H_\theta : x \mapsto \nabla_\theta \log(p(x, y|\theta))$ and π_θ is the probability distribution with density with respect to the Lebesgue measure $x \mapsto p(x|y, \theta)$.

To minimize the objective function f we suggest the use of a SA strategy which extends the one presented in Section 2. More precisely, motivated by the methodology described in Section 2, we propose a SA scheme which relies on biased estimates of $\nabla f(\theta)$ through a family of Markov kernels $\{\mathrm{K}_{\gamma,\theta}, \gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta\}$, for $\bar{\gamma} > 0$, such that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, $\mathrm{K}_{\gamma,\theta}$ admits an invariant probability distribution $\pi_{\gamma,\theta}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In the SOUL method, the Markov kernel $\mathrm{K}_{\gamma,\theta}$ stands for $\mathrm{R}_{\gamma,\theta}$ for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, where $\mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with (6). We assume in addition that the bias associated to the use of this family of Markov kernels can be controlled with respect to to γ uniformly in θ, *i.e.* for example there exists $C > 0$ such that for all $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, $\|\pi_{\gamma,\theta} - \pi_\theta\|_{\mathrm{TV}} \leq C\gamma^\alpha$ with $\alpha > 0$.

Let now $(\delta_n)_{n \in \mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ and $(m_n)_{n \in \mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$ be sequences of stepsizes and batch sizes which will be used to define the sequence relatively to the variable θ similarly to (2) and (3). Let $(\gamma_n)_{n \in \mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ be a sequence of stepsizes which will be used to get approximate samples from π_{θ_n}, similarly to (6). Starting from $X_0^0 \in \mathbb{R}^d$ and $\theta_0 \in \Theta$, we define on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

$(\{X_k^n : k \in \{0,\dots,m_n\}\}, \theta_n)_{n\in\mathbb{N}}$ by the following recursion for $n \in \mathbb{N}$ and $k \in \{0,\dots,m_n - 1\}$

$$
\begin{aligned}
&(X_k^n)_{k\in\{0,\dots,m_n\}} \text{ is a MC with kernel } \mathrm{K}_{\gamma_n,\theta_n} \text{ and } X_0^n = X_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1} , \\
&\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n) \right] ,
\end{aligned}
\tag{14}
$$

where Π_Θ is the projection onto Θ and $\mathcal{F}_n$ is defined as follows for all $n \in \mathbb{N}$

$$
\mathcal{F}_n = \sigma\left(\theta_0, \{(X_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}\right) , \qquad \mathcal{F}_{-1} = \sigma(\theta_0, X_0^0) \tag{15}
$$

where $\{(X_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}$ is given by (14). Note that such a construction is always possible by the Kolmogorov extension theorem [34, Theorem 5.16], and by (14), for any $n \in \mathbb{N}$, θ_{n+1} is $\mathcal{F}_n$-measurable. Then the sequence of approximate minimizers of f is given by $(\hat{\theta}_N)_{N\in\mathbb{N}}$, (8).

Under different sets of conditions on $f, H, (\delta_n)_{n\in\mathbb{N}}, (\gamma_n)_{n\in\mathbb{N}}$ and $(m_n)_{n\in\mathbb{N}}$ we obtain that $(\theta_n)_{n\in\mathbb{N}}$ converges almost surely to an element of $\arg\min_\Theta f$. In particular in this section we consider the case where f is assumed to be convex. We establish that if $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ go to 0 sufficiently fast, $\mathbb{E}[f(\hat{\theta}_N)] - \min_\Theta f$ goes to 0 with a quantitative rate of convergence. In the case where $(\gamma_n)_{n\in\mathbb{N}}$ is held fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma$, we show that while $\mathbb{E}[f(\hat{\theta}_N)]$ does not converge to 0, there exists $C, \alpha > 0$ such that $\limsup_{N\to+\infty} \mathbb{E}[f(\hat{\theta}_N)] - \min_\Theta f \leq C\gamma^\alpha$. In the case where f is non-convex, we apply some results from stochastic approximation [37] which establish that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges almost surely to a stationary point of the projected ordinary differential equation associated with ∇f and Θ. We postpone this result to Appendix B, since it involves a theoretical background which we think is out of the scope of the main document.

4.3 Main results

We impose a stability condition on the stochastic process $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ defined by (14) and that for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$ the iterates of $\mathrm{K}_{\gamma,\theta}$ are close enough to π_θ after a sufficiently large number of iterations.

H1. *There exists a measurable function $V : \mathbb{R}^d \to [1, +\infty)$ satisfying the following conditions.*

(i) There exists $A_1 \geq 1$ such that for any $n, p \in \mathbb{N}$, $k \in \{0,\dots,m_n\}$

$$
\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^p V(X_k^n) \middle| X_0^0 \right] \leq A_1 V(X_0^0) , \qquad \mathbb{E}\left[V(X_0^0)\right] < +\infty ,
$$

where $\{(X_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}$ is given by (14).

(ii) There exist $A_2, A_3 \geq 1$, $\rho \in [0,1)$ such that for any $\gamma \in (0,\bar{\gamma}]$, $\theta \in \Theta$, $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$, $\mathrm{K}_{\gamma,\theta}$ has a stationary distribution $\pi_{\gamma,\theta}$ and

$$
\|\updelta_x \mathrm{K}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_V \leq A_2 \rho^{n\gamma} V(x) , \qquad \pi_{\gamma,\theta}(V) \leq A_3 .
$$

(iii) There exists $\mathbf{\Psi} : \mathbb{R}_+^\star \to \mathbb{R}_+$ such that for any $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \Theta$

$$
\|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \leq \mathbf{\Psi}(\gamma) .
$$

H1-(ii) is an ergodicity condition in V-norm for the Markov kernel $\mathrm{K}_{\gamma,\theta}$ uniform in $\theta \in \Theta$. There exists an extensive literature on the conditions under which a Markov kernel is ergodic [41, 19]. **H** 1-(iii) ensures that the distance between the invariant measure $\pi_{\gamma,\theta}$ of the Markov kernel $\mathrm{K}_{\gamma,\theta}$ and π_θ can be controlled uniformly in θ. We show that this condition holds in the case of the Langevin Monte Carlo algorithm in Proposition 23.

We now state our mains results.

Theorem 1 (Increasing batch size 1)**.** *Assume* **A**1*,* **A**2*,* **A**3 *hold and f is convex. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be sequences of positive integers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/L_f$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$ and*

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1} \boldsymbol{\Psi}(\gamma_n) < +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}/(m_n \gamma_n) < +\infty \;. \tag{16}$$

Let $\{(X_k^n)_{k\in\{0,\ldots,m_n\}} : n \in \mathbb{N}\}$ and $(\theta_n)_{n\in\mathbb{N}}$ be given by (14). Assume in addition that **H**1 *is satisfied and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/2}(x)$. Then, the following statements hold:*

(a) $(\theta_n)_{n\in\mathbb{N}}$ converges almost surely to some $\theta^\star \in \arg\min_\Theta f$;

(b) furthermore, almost surely there exists $C \ge 0$ such that for any $n \in \mathbb{N}^*$

$$\left\{ \sum_{k=1}^{n} \delta_k f(\theta_k) \middle/ \sum_{k=1}^{n} \delta_k \right\} - \min_\Theta f \le C \middle/ \left(\sum_{k=1}^{n} \delta_k \right) \;.$$

Proof. The proof is postponed to Appendix C.1. □

Note that in (14), $X_0^n = X_{m_{n-1}}^{n-1}$ for $n \in \mathbb{N}^*$. This procedure is referred to as warm-start in the sequel. An inspection of the proof of Theorem 1 shows that X_0^n could be any random variable independent from $\mathcal{F}_{n-1}$ for any $n \in \mathbb{N}$ with $\sup_{n\in\mathbb{N}^*} \mathbb{E}[V(X_0^n)] < +\infty$. It is not an option in the fixed batch size setting of Theorem 3, where the warm-start procedure is crucial for the convergence to occur.

We extend this theorem to non convex objective function see Theorem 8 in Appendix B. Under the conditions of Theorem 1 with the additional assumption that $\partial\Theta$ is a smooth manifold we obtain that $(\theta_n)_{n\in\mathbb{N}}$ converges almost surely to some point θ^* such that $\nabla f(\theta^*) + \mathbf{n} = 0$ with $\mathbf{n} = 0$ if $\theta^* \in \mathrm{int}(\Theta)$ and $\mathbf{n} \in \mathrm{T}(\theta^*, \partial\Theta)^\perp$ if $\theta^* \in \partial\Theta$, where $\mathrm{T}(\theta, \partial\Theta)$ is the tangent space of $\partial\Theta$ at point $\theta \in \partial\Theta$, see [3, Chapter 2].

In the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin update (6), under appropriate conditions Proposition 23 shows that for any $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, $\boldsymbol{\Psi}(\gamma) = \mathcal{O}(\gamma^{1/2})$. In that case, assume then that there exist $a, b, c > 0$ such that for any $n \in \mathbb{N}^*$, $\delta_n = n^{-a}$, $\gamma_n = n^{-b}$ and $m_n = \lceil n^c \rceil$ then (16) is equivalent to

$$a < 1 \;, \qquad a + b/2 > 1 \;, \qquad a - b + c > 1 \;. \tag{17}$$

Suppose $a \in [0, 1)$ is given, then the previous equation reads

$$b = 2(1-a) + \varsigma_1 \;, \qquad c = 3(1-a) + \varsigma_2 \;, \qquad \varsigma_2 > \varsigma_1 > 0 \;.$$

This illustrates a trade-off between the intrinsic inaccuracy of our algorithm through the family of Markov kernels (14) which do not exactly target π_θ and the minimization aim of our scheme. Note also that $(\delta_n)_{n\in\mathbb{N}}$ is allowed to be constant. This case yields $\gamma_n = n^{-2-\varsigma_1}$ and $m_n = \lceil n^{3+\varsigma_2} \rceil$ with $\varsigma_2 > \varsigma_1 > 0$.

In our next result we derive an non-asymptotic upper-bound of $(\mathbb{E}[f(\hat{\theta}_n) - \min_\Theta f])_{n\in\mathbb{N}}$.

Theorem 2 (Increasing batch size 2)**.** *Assume* **A**1*,* **A**2*,* **A**3 *hold and f is convex. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of positive integers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/L_f$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$. Let $\{(X_k^n)_{k\in\{0,\ldots,m_n\}} : n \in \mathbb{N}\}$ be given by (14). Assume in addition that* **H**1 *is satisfied and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. Then, there exists $(E_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^*$*

$$\mathbb{E}\left[\left\{\sum_{k=1}^{n} \delta_k f(\theta_k) \Big/ \sum_{k=1}^{n} \delta_k\right\} - \min_\Theta f\right] \leq E_n \Big/ \left(\sum_{k=1}^{n} \delta_k\right) ,$$

with for any $n \in \mathbb{N}^*$,

$$\begin{aligned} E_n = 2M_\Theta^2 + 2B_1 M_\Theta \mathbb{E}\left[V^{1/2}(X_0^0)\right] \sum_{k=0}^{n-1} \delta_{k+1}/(m_k\gamma_k) \\ + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1}\mathbf{\Psi}(\gamma_k) + 4B_1^2 \mathbb{E}\left[V(X_0^0)\right] \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 \\ + 4\sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbf{\Psi}(\gamma_k)^2 + B_2 \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 , \end{aligned} \tag{18}$$

where B_1 and B_2 are given in Lemma 11 and Lemma 12 respectively.

Proof. The proof is postponed to Appendix C.2. □

We recall that in the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin update (6), under appropriate conditions Proposition 23 shows that for any $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, $\mathbf{\Psi}(\gamma) = \mathcal{O}(\gamma^{1/2})$. In that case, if there exist $a, b, c \geq 0$ such that for any $n \in \mathbb{N}^*$, $\delta_n = n^{-a}$, $\gamma_n = n^{-b}$, $m_n = n^c$ and (17) holds, the accuracy, respectively the complexity, of the algorithm are of orders $\left(\sum_{k=1}^n \delta_k\right)^{-1} = \mathcal{O}(n^{a-1})$, respectively $\sum_{k=0}^n m_k = \mathcal{O}(n^{3(1-a)+\varsigma_2+1})$ for $\varsigma_2 > 0$. Thus, for a fix target precision $\varepsilon > 0$, it requires that $\varepsilon = \mathcal{O}(n^{a-1})$ and the complexity reads $\mathcal{O}(\varepsilon^{-3} \left(\log(1/\varepsilon)/(1-a)\right)^{1+\varsigma_2})$. On the other hand, if we fix the complexity budget to N the accuracy is of order $\mathcal{O}(N^{-(3+(1+\varsigma_2)/(1-a))^{-1}})$. These two considerations suggest to set a close to 0. In the special case where $a = 0$, we obtain that the accuracy is of order $\mathcal{O}(n^{-1})$, which is similar to the order identified in the deterministic gradient descent for convex functionals.

A case of interest is the fix stepsize setting, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma > 0$. Assume that $(\delta_n)_{n\in\mathbb{N}}$ is non-increasing $\lim_{n\to+\infty} \delta_n = 0$ and $\lim_{n\to+\infty} m_n = +\infty$. In addition, assume that $\sum_{n\in\mathbb{N}^*} \delta_n = +\infty$ then, by [46, Problem 80, Part I], it holds that

$$\begin{cases} \lim_{n\to+\infty} \left[(\sum_{k=1}^n \delta_k/m_k)/(\sum_{k=1}^n \delta_k)\right] = \lim_{n\to+\infty} 1/m_n = 0 \, ; \\ \lim_{n\to+\infty} \left[\left(\sum_{k=1}^n \delta_k^2\right)/(\sum_{k=1}^n \delta_k)\right] = \lim_{n\to+\infty} \delta_n = 0 \, . \end{cases}$$

Therefore, we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{\sum_{k=1}^{n}\delta_k f(\theta_k)\Bigg/\sum_{k=1}^{n}\delta_k\right\} - \min f\right] \leq 2M_\Theta \boldsymbol{\Psi}(\gamma) \;.$$

Similarly, if the stepsize is fixed and the number of Markov chain iterates is fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma$ and $m_n = m$ with $\gamma > 0$ and $m \in \mathbb{N}^*$, we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{\sum_{k=1}^{n}\delta_k f(\theta_k)\Bigg/\sum_{k=1}^{n}\delta_k\right\} - \min f\right] \leq \boldsymbol{\Xi}_1(\gamma) \;, \tag{19}$$

with

$$\boldsymbol{\Xi}_1(\gamma) = 2B_1 M_\Theta \mathbb{E}\left[V^{1/2}(X_0^0)\right]/\gamma + 2M_\Theta \boldsymbol{\Psi}(\gamma) \;.$$

However if $(m_n)_{n\in\mathbb{N}}$ is constant the convergence cannot be obtained using Theorem 1. Strengthening the conditions of Theorem 1 and making use of the warm-start property of the algorithm we can derive the convergence in that case.

We now are interested in the case where the batch size is fixed, *i.e.* $m_n = m_0$ for all $n \in \mathbb{N}$. For ease of exposition we only consider $m_0 = 1$ and let $\tilde{X}_{n+1} = X_1^n$ for any $n \in \mathbb{N}$. However the general case can be adapted from the proof of the result stated below. More precisely we consider the setting where the recursion (14) can be written for any $n \in \mathbb{N}$ as

$$\begin{aligned} &\tilde{X}_{n+1} \text{ has distribution } \mathrm{K}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n,\cdot) \text{ conditionally to } \tilde{\mathcal{F}}_n \;, \\ &\tilde{\theta}_{n+1} = \Pi_\Theta\left[\tilde{\theta}_n - \delta_{n+1} H_{\tilde{\theta}_n}(\tilde{X}_{n+1})\right] \;, \end{aligned} \tag{20}$$

with $\theta_0 \in \Theta$, $\tilde{X}_0 \in \mathbb{R}^d$ and where $\tilde{\mathcal{F}}_n$ is given by

$$\tilde{\mathcal{F}}_n = \sigma\left(\tilde{\theta}_0, (\tilde{X}_\ell)_{\ell\in\{0,\ldots,n\}}\right) \;. \tag{21}$$

We consider the following assumption on the family $\{H_\theta \;:\; \theta \in \Theta\}$.

A4. *There exists $L_H \geq 0$ such that for any $x \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in \Theta$,*

$$\|H_{\theta_1}(x) - H_{\theta_2}(x)\| \leq L_H \|\theta_1 - \theta_2\| V^{1/2}(x) \;.$$

We consider a similar property as **A**4 on the family of Markov kernels $\{\mathrm{K}_{\gamma,\theta}, \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$, which weakens the assumption [2, H6].

H2. *There exist a measurable function $V : \mathbb{R}^d \to [1,+\infty)$, $\boldsymbol{\Lambda}_1 : \left(\mathbb{R}_+^*\right)^2 \to \mathbb{R}_+$ and $\boldsymbol{\Lambda}_2 : \left(\mathbb{R}_+^*\right)^2 \to \mathbb{R}_+$ such that for any $\gamma_1, \gamma_2 \in (0,\bar{\gamma}]$ with $\gamma_2 < \gamma_1$, $\theta_1, \theta_2 \in \Theta$, $x \in \mathbb{R}^d$ and $a \in [1/4, 1/2]$*

$$\|\delta_x \mathrm{K}_{\gamma_1,\theta_1} - \delta_x \mathrm{K}_{\gamma_2,\theta_2}\|_{V^a} \leq \left[\boldsymbol{\Lambda}_1(\gamma_1,\gamma_2) + \boldsymbol{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right] V^{2a}(x) \;.$$

The following theorem ensures convergence properties for $(\theta_n)_{n\in\mathbb{N}}$ similar to the ones of Theorem 1. The proof of this result is based on a generalization of [30, Lemma 4.2] for inexact MCMC schemes.

Theorem 3 (Fixed batch size 1)**.** *Assume* **A**1*,* **A**2*,* **A**3*,* **A**4 *hold and f is convex. Let $\bar{\gamma} > 0$, $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/L_f$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n|\delta_n^{-2} < +\infty$, $\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty$ and*

$$\begin{aligned} &\sum_{n=0}^{+\infty} \delta_{n+1}\boldsymbol{\Psi}(\gamma_n) < +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}^2\gamma_n^{-2} < +\infty \;, \\ &\sum_{n=0}^{+\infty} \delta_{n+1}\gamma_{n+1}^{-2}\left[\boldsymbol{\Lambda}_1(\gamma_n,\gamma_{n+1}) + \delta_{n+1}\boldsymbol{\Lambda}_2(\gamma_n,\gamma_{n+1})\right] < +\infty \;. \end{aligned} \tag{22}$$

Let $(\tilde{X}_n)_{n\in\mathbb{N}}$ be given by (20). Assume in addition that **H**1 *and* **H**2 *are satisfied and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$. Then the following statements hold:*

(a) $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ converges almost surely to some $\theta^\star \in \arg\min_\Theta f$;

(b) furthermore, almost surely there exists $C \ge 0$ such that for any $n \in \mathbb{N}^*$

$$\left\{ \sum_{k=1}^n \delta_k f(\tilde{\theta}_k) \middle/ \sum_{k=1}^n \delta_k \right\} - \min_\Theta f \le C \middle/ \left(\sum_{k=1}^n \delta_k \right) \;.$$

Proof. The proof is postponed to Appendix C.3. □

In the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin update (6), under appropriate conditions Proposition 23 and Proposition 24 show that for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and $\gamma_1 > \gamma_2$, $\boldsymbol{\Psi}(\gamma_1) = \mathrm{C}_1\gamma^{1/2}$, $\boldsymbol{\Lambda}_1(\gamma_1,\gamma_2) = \mathrm{C}_2(\gamma_1/\gamma_2 - 1)$ and $\boldsymbol{\Lambda}_2(\gamma_1,\gamma_2) = \mathrm{C}_3\gamma_2^{1/2}$, for $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 \ge 0$. Thus we obtain that the following series should converge

$$\begin{aligned} &\sum_{n=0}^{+\infty} \delta_{n+1}\gamma_n^{1/2} < +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}^2/\gamma_{n+1}^2 < +\infty \;, \\ &\sum_{n=0}^{+\infty} \delta_{n+1}(\gamma_n - \gamma_{n+1})/\gamma_{n+1}^3 < +\infty \;. \end{aligned} \tag{23}$$

If there exist $a, b > 0$ such that $\delta_n = n^{-a}$ and $\gamma_n = n^{-b}$, then (23) is satisfied if $b \in (2(1-a), a - 1/2)$ which is not empty if $a > 5/6$.

Theorem 4 (Fixed batch size 2)**.** *Assume* **A**1*,* **A**2*,* **A**3*,* **A**4 *hold and f is convex. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of positive integers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/L_f$ and $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$. Let $(\tilde{X}_n)_{n\in\mathbb{N}}$ be given by (20). Assume in addition that* **H**1 *and* **H**2 *are satisfied and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$. Then, there exists $(\tilde{E}_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^*$*

$$\mathbb{E}\left[\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \middle/ \sum_{k=1}^n \delta_k \right\} - \min_\Theta f \right] \le \tilde{E}_n \middle/ \left(\sum_{k=1}^n \delta_k \right) \;,$$

with for any $n \in \mathbb{N}^*$,

$$\begin{aligned}\tilde{E}_n &= 2M_\Theta + 2M_\Theta \sum_{k=0}^{n} \delta_{k+1}\boldsymbol{\Psi}(\gamma_k) + C_3 \sum_{k=0}^{n} |\delta_{k+1} - \delta_k| \gamma_k^{-1} \\ &+ 2M_\Theta C_2 \sum_{k=0}^{n} \delta_{k+1}\gamma_{k+1}^{-1} \left[\gamma_{k+1}^{-1} \{\boldsymbol{\Lambda}_1(\gamma_k, \gamma_{k+1}) + \boldsymbol{\Lambda}_2(\gamma_k, \gamma_{k+1})\delta_{k+1}\} + \delta_{k+1}\right] \\ &+ C_3 \sum_{k=0}^{n} \delta_{k+1}^2 \gamma_{k+1}^{-1} + C_3(\delta_{n+1}/\gamma_n - \delta_0/\gamma_0) + C_1 \sum_{k=0}^{n} \delta_{k+1}^2 \ .\end{aligned}$$

where C_1, C_2 and C_3 are given in Lemma 13, Lemma 16 and Lemma 15 respectively.

Proof. The proof is postponed to Appendix C.4. □

Theorem 4 improves the conclusions of Theorem 2 in the case where $\gamma_n = \gamma > 0$ for any $n \in \mathbb{N}$. Indeed, in that case, similarly to (19), assuming that $\lim_{n\to+\infty} \delta_n = 0$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n| \delta_n^{-2} < +\infty$, $\boldsymbol{\Lambda}_1(t,t) = 0$ for any $t > 0$, we obtain that for all $n \in \mathbb{N}$

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{\sum_{k=1}^{n} \delta_k f(\theta_k) \middle/ \sum_{k=1}^{n} \delta_k\right\} - \min f\right] \leq \boldsymbol{\Xi}_2(\gamma) \ ,$$

with $\boldsymbol{\Xi}_2(\gamma) = 2M_\Theta \Psi(\gamma) \leq \boldsymbol{\Xi}_1(\gamma) = 2B_1 M_\Theta \mathbb{E}\left[V^{1/2}(X_0^0)\right]/\gamma + 2M_\Theta \boldsymbol{\Psi}(\gamma)$. In the case where $\sup_{\gamma\in(0,\bar{\gamma}]} \boldsymbol{\Psi}(\gamma) < +\infty$, $\boldsymbol{\Xi}_2(\gamma)$ is of order $\mathcal{O}(\boldsymbol{\Psi}(\gamma))$ and $\boldsymbol{\Xi}_1(\gamma)$ is of order $\mathcal{O}(\gamma^{-1})$. Therefore if $\lim_{\gamma\to 0} \boldsymbol{\Psi}(\gamma) = 0$, even in the fixed batch size setting, the minimum of the objective function f can be approached with arbitrary precision $\varepsilon > 0$ by choosing γ small enough.

4.4 Application to SOUL

We now apply our results to the SOUL methodology introduced in Section 2 where the Markov kernel $\mathrm{R}_{\gamma,\theta}$ with $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$ is given by a Langevin Markov kernel and associated with recursion (6). Setting for any $\theta \in \Theta$, $\pi_\theta = p(\cdot|y, \theta)$, we consider the following assumption on the family of probability distributions $(\pi_\theta)_{\theta\in\Theta}$.

L1. *For any $\theta \in \Theta$, there exists $U_\theta : \mathbb{R}^d \to \mathbb{R}$ such that π_θ admits a probability density function with respect to to the Lebesgue measure proportional to $x \mapsto \exp(-U_\theta(x))$. In addition $(\theta, x) \mapsto U_\theta(x)$ is continuous, $x \mapsto U_\theta(x)$ is differentiable for all $\theta \in \Theta$ and there exists $\mathtt{L} \geq 0$ such that for any $x, y \in \mathbb{R}^d$,*

$$\sup_{\theta\in\Theta} \|\nabla_x U_\theta(x) - \nabla_x U_\theta(y)\| \leq \mathtt{L} \|x - y\| \ ,$$

and $\{\|\nabla_x U_\theta(0)\| : \theta \in \Theta\}$ is bounded.

In the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, the first line of (14) can be rewritten for any $n \in \mathbb{N}$ and $k \in \{0, \ldots, m_n - 1\}$

$$X_{k+1}^n = X_k^n - \gamma_n \nabla_x U_{\theta_n}(X_k^n) + \sqrt{2\gamma_n} Z_{k+1}^n \ , \text{ with } X_0^n = X_{m_{n-1}}^{n-1} \text{ if } n \geq 1 \ , \tag{24}$$

given $(\gamma_n)_{n\in\mathbb{N}} \in (0, \bar{\gamma}]^{\mathbb{N}}$, $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$ and $(Z_k^n)_{n\in\mathbb{N}, k\in\{1,\ldots,m_n\}}$ a family of i.i.d d-dimensional zero-mean Gaussian random variables with covariance matrix identity. In the following propositions, we show that the results above hold by deriving sufficient conditions under which **H**1 and **H**2 are satisfied.

Under **L**1, the Langevin diffusion defined by (5) admits a unique strong solution for any $\theta \in \Theta$. Consider now the following additional tail condition on U_θ which ensures geometric ergodicity of $\mathrm{R}_{\gamma,\theta}$ for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, with $\bar{\gamma}$ which will be specified below.

L2. *There exist* $\mathtt{m}_1 > 0$ *and* $\mathtt{m}_2, \mathtt{c}, R_1 \geq 0$ *such that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$,

$$\langle \nabla_x U_\theta(x), x \rangle \geq \mathtt{m}_1 \|x\| \mathbb{1}_{\mathrm{B}(0,R_1)^c}(x) + \mathtt{m}_2 \|\nabla_x U_\theta(x)\|^2 - \mathtt{c} \ .$$

L3. *There exists* $L_U \geq 0$ *such that for any* $x \in \mathbb{R}^d$ *and* $\theta_1, \theta_2 \in \Theta$

$$\|\nabla_x U_{\theta_1}(x) - \nabla_x U_{\theta_2}(x)\| \leq L_U \|\theta_1 - \theta_2\| V(x)^{1/2} \ .$$

The next theorems assert that under **L**1, **L**2 and **L**3 the SOUL algorithm introduced in Section 2 satisfy **H**1 and **H**2 and therefore Theorem 1, Theorem 2, Theorem 3 and Theorem 4 can be applied if in addition **A**1, **A**2, **A**3 and **A**4 hold.

Under **L**2 define for any $x \in \mathbb{R}^d$

$$V_{\mathrm{e}}(x) = \exp\left[\mathtt{m}_1 \sqrt{1 + \|x\|^2}/4\right] \ .$$

Theorem 5. *Assume* **L**1 *and* **L**2*. Then,* **H**1 *holds with* $V \leftarrow V_{\mathrm{e}}$, $\bar{\gamma} \leftarrow \min(1, 2\mathtt{m}_2)$ *and* $\boldsymbol{\Psi}(\gamma) = D_4\sqrt{\gamma}$ *where* D_4 *is given in Proposition 23.*

Proof. The proof is postponed to Appendix C.5. □

Theorem 6. *Assume* **L**1, **L**2, **L**3 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V_{\mathrm{e}}^{1/4}(x)$. **H**2 *holds with* $V \leftarrow V_{\mathrm{e}}$ *and* $\bar{\gamma} \leftarrow \min(1, 2\mathtt{m}_2)$ *and for any* $\gamma_1, \gamma \in (0, \bar{\gamma}]$, $\gamma_2 < \gamma_1$,

$$\boldsymbol{\Lambda}_1(\gamma_1, \gamma_2) = D_5(\gamma_1/\gamma_2 - 1) \ , \quad \boldsymbol{\Lambda}_2(\gamma_1, \gamma_2) = D_5 \gamma_2^{1/2} \ ,$$

where D_5 *is given in Proposition 24 in Appendix C.6.*

Proof. The proof is postponed to Appendix C.6. □

References

[1] C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006. ISSN 1050-5164. doi: 10.1214/105051606000000286. URL https://doi.org/10.1214/105051606000000286.

[2] Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res*, 18(1):310–342, 2017.

[3] T. Aubin. *A course in differential geometry*. Graduate Studies in Mathematics. AMS, 2000. ISBN 9780821827093,082182709X.

[4] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014. ISBN 978-3-319-00226-2; 978-3-319-00227-9. doi: 10.1007/978-3-319-00227-9. URL http://dx.doi.org/10.1007/978-3-319-00227-9.

[5] Laura Balzano, Robert Nowak, and J Ellenberg. Compressed sensing audio demonstration. *website http://web.eecs.umich.edu/~girasole/csaudio*, 2010.

[6] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. ISBN 3-540-52894-6. doi: 10.1007/978-3-642-75894-2. URL https://doi.org/10.1007/978-3-642-75894-2. Translated from the French by Stephen S. Wilson.

[7] R. Berger and G. Casella. *Statistical inference (2nd ed.)*. Duxbury / Thomson Learning, Pacific Grove, USA, 2002.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[10] Emmanuel J Candès et al. Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.

[11] Bradley P. Carlin and Thomas A. Louis. Empirical Bayes: past, present and future. *J. Amer. Statist. Assoc.*, 95(452):1286–1289, 2000. ISSN 0162-1459. doi: 10.2307/2669771. URL https://doi.org/10.2307/2669771.

[12] G. Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

[13] George Casella. An introduction to empirical Bayes data analysis. *Amer. Statist.*, 39(2):83–87, 1985. ISSN 0003-1305. doi: 10.2307/2682801. URL https://doi.org/10.2307/2682801.

[14] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[15] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017. ISSN 1369-7412. doi: 10.1111/rssb.12183. URL https://doi.org/10.1111/rssb.12183.

[16] V. De Bortoli and A. Durmus. Convergence of diffusions and their discretizations:from continuous to discrete processes and back. *arXiv preprint arXiv:1904.09808*, 2019.

[17] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. URL https://doi.org/10.1214/aos/1018031103.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B*, 39(1):1–38, 1977.

[19] R. Douc, É. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer, 2018. to be published.

[20] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. ISSN 1532-4435.

[21] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017. ISSN 1050-5164. doi: 10.1214/16-AAP1238. URL https://doi.org/10.1214/16-AAP1238.

[22] A. Durmus, E. Moulines, and E. Saksman. On the convergence of hamiltonian monte carlo. *arXiv preprint arXiv:1705.00166*, 2017.

[23] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[24] A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

[25] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, 166(3-4):851–886, 2016. ISSN 0178-8051. doi: 10.1007/s00440-015-0673-1. URL https://doi.org/10.1007/s00440-015-0673-1.

[26] Andreas Eberle and Mateusz B Majka. Quantitative contraction rates for markov chains on general state spaces. *arXiv preprint arXiv:1808.07033*, 2018.

[27] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. ISBN 0-471-08186-8. doi: 10.1002/9780470316658. URL https://doi.org/10.1002/9780470316658. Characterization and convergence.

[28] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007.

[29] G. Fort and E. Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259, 2003. ISSN 0090-5364. doi: 10.1214/aos/1059655912. URL https://doi.org/10.1214/aos/1059655912.

[30] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS938. URL https://doi.org/10.1214/11-AOS938.

[31] R. J. Gardner. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc. (N.S.)*, 39(3):355–405, 2002. ISSN 0273-0979. doi: 10.1090/S0273-0979-02-00941-2. URL https://doi.org/10.1090/S0273-0979-02-00941-2.

[32] James E Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of computational statistics: concepts and methods*. Springer Science & Business Media, 2012.

[33] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011. doi: 10.1111/j.1467-9868.2010.00765.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00765.x.

[34] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[35] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. A method for large-scale l1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 1(4):606–617, 2007.

[36] Solomon Kullback. *Information theory and statistics*. Dover Publications, Inc., Mineola, NY, 1997. ISBN 0-486-69684-7. Reprint of the second (1968) edition.

[37] Harold J. Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. ISBN 0-387-00894-2. Stochastic Modelling and Applied Probability.

[38] Sajan Goud Lingala and Mathews Jacob. A blind compressive sensing frame work for accelerated dynamic mri. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1060–1063. IEEE, 2012.

[39] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993. ISBN 3-540-19832-6. doi: 10.1007/978-1-4471-3267-7. URL https://doi.org/10.1007/978-1-4471-3267-7.

[40] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.*, 25(3):518–548, 1993. ISSN 0001-8678. doi: 10.2307/1427522. URL http://dx.doi.org/10.2307/1427522.

[41] Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes. I. Criteria for discrete-time chains. *Adv. in Appl. Probab.*, 24(3):542–574, 1992. ISSN 0001-8678. doi: 10.2307/1427479. URL https://doi.org/10.2307/1427479.

[42] Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.*, 25(3):518–548, 1993. ISSN 0001-8678. doi: 10.2307/1427522. URL https://doi.org/10.2307/1427522.

[43] Vishal Monga. *Handbook of Convex Optimization Methods in Imaging Science*. Springer, 2017.

[44] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008. ISSN 1052-6234. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

[45] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504): 1339–1349, 2013.

[46] George Pólya and Gabor Szegő. *Problems and theorems in analysis. I*. Classics in Mathematics. Springer-Verlag, Berlin, 1998. ISBN 3-540-63640-4. doi: 10.1007/978-3-642-61905-2. URL https://doi.org/10.1007/978-3-642-61905-2. Series, integral calculus, theory of

functions, Translated from the German by Dorothee Aeppli, Reprint of the 1978 English translation.

[47] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999. ISBN 3-540-64325-7. doi: 10.1007/978-3-662-06400-9. URL https://doi.org/10.1007/978-3-662-06400-9.

[48] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

[49] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd ed.)*. Springer-Verlag, New York, 2004.

[50] Christian P Robert and Darren Wraith. Computational methods for bayesian model choice. In *Aip conference proceedings*, volume 1193, pages 251–262. AIP, 2009.

[51] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 1350-7265. URL https://doi.org/10.2307/3318418.

[52] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. ISBN 978-3-540-28998-2; 3-540-28998-4. Reprint of the 1997 edition.

[53] A. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus. Maximum likelihood estimation of regularisation parameters: An empirical bayesian approach. 2019. in prepration.

[54] Ana Fernandez Vidal and Marcelo Pereyra. Maximum likelihood estimation of regularisation parameters. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1742–1746. IEEE, 2018.

[55] Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Science & Business Media, 2013.

[56] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X; 0-521-40605-6. doi: 10.1017/CBO9780511813658. URL https://doi.org/10.1017/CBO9780511813658.

A Posterior convexity

Lemma 7. *For any $y \in \{0,1\}^{d_y}$, $\theta \mapsto p(y|\theta)$ given by (9) is log-concave.*

Proof. Let $\theta \in \mathbb{R}$, then by (9), for any $y \in \mathbb{R}$ we have $p(y|\theta) = \int_{\mathbb{R}^d} p(y, \beta|\theta)\mathrm{d}\beta$ with

$$p(y, \beta|\theta) = (2\pi\sigma^2)^{-d/2} \left\{ \prod_{i=1}^{d_y} s(x_i^{\mathrm{T}}\beta)^{y_i}(1 - s(x_i^{\mathrm{T}}\beta))^{1-y_i} \right\} \mathrm{e}^{-\frac{\|\beta - \theta \mathbf{1}_d\|^2}{2\sigma^2}} \ .$$

Therefore we have using that for any $t \in \mathbb{R}$, $1 - s(t) = s(-t)$

$$\log p(y, \beta|\theta) = (-d/2)\log(2\pi\sigma^2) + \left\{ \sum_{i=1}^{d_y} y_i \log(s(x_i^{\mathrm{T}}\beta)) + (1 - y_i)\log(s(-x_i^{\mathrm{T}}\beta)) \right\} - \frac{\|\beta - \theta \mathbf{1}_d\|^2}{2\sigma^2} \, .$$

Since $y_i \geq 0$, $1 - y_i \geq 0$, $(\beta, \theta) \mapsto \|\beta - \theta\mathbf{1}_d\|^2$, $t \mapsto \log(s(t))$ and $t \mapsto \log(s(-t))$ are convex, we obtain that $(\beta, \theta) \mapsto p(y, \beta|\theta)$ is log-concave. Using the Prékopa–Leindler inequality [31, Theorem 7.1] we obtain that $\theta \mapsto p(y|\theta)$ is log-concave which concludes the proof. □

B Non-convex objective function

In this section we turn to the case where f is non-convex. We recall that the normal space of a sub-manifold $\mathcal{M} \subset \mathbb{R}^{d_\Theta}$ at point θ is given by

$$\mathrm{N}(\theta, \mathcal{M}) = \begin{cases} \mathrm{T}(\theta, \mathcal{M})^{\perp} & \text{if } \theta \in \mathcal{M} \, ; \\ \{0\} & \text{otherwise} \, , \end{cases}$$

where $\mathrm{T}(\theta, \mathcal{M})$ is the tangent space of the sub-manifold $\mathcal{M}$ at point x, see [3].

Theorem 8. *Assume* **A**1*,* **A**2*,* **A**3 *and that Θ is a d_Θ dimensional connected differentiable manifold with boundary and continuously differentiable outer normal. Let $\bar{\gamma} > 0$, $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of positive integers such that $\sup_{n\in\mathbb{N}} \delta_n < 1/L_f$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$ and (16) are satisfied. Let $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ be given by (14). Assume in addition that* **H**1 *is satisfied. Then $(\theta_n)_{n\in\mathbb{N}}$ defined by (14) converges almost surely to some $\theta^\star \in \{\theta \in \Theta : \nabla f(\theta) + \mathbf{n} = 0, \; \mathbf{n} \in \mathrm{N}(\theta, \partial\Theta)\}$.*

Proof. The proof is an application of [37, Chapter 5, Theorem 2.3] using the decomposition of the error term considered in the proof of Theorem 1 and Theorem 3. Indeed we decompose the error term η_n defined by (25) as $\eta_n = \delta M_n + B_n$, where δM_n is a martingale increment. Then, we only need to show that the following sums converge

$$\sum_{k=0}^{n} \delta_{k+1}^2 \mathbb{E}\left[\|\delta M_k\|^2\right] \, , \qquad \sum_{k=0}^{n} \delta_{k+1} \mathbb{E}\left[\|B_k\|\right] \, ,$$

which is established in Lemma 11 and Lemma 12. □

C Postponed proofs

We first derive the following technical lemmas.

Lemma 9. *Let $t \in (0, 1)$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ then $\sum_{n\in\mathbb{N}} t^{n\gamma} \leq t^{-\bar{\gamma}} \log^{-1}(1/t)\gamma^{-1}$ and $\sum_{n\in\mathbb{N}} n t^{n\gamma} \leq t^{-\bar{\gamma}} \log^{-2}(1/t)\gamma^{-2}$.*

Proof. Let $t \in (0,1)$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$. Using that $e^u - 1 \leq ue^u$ for all $u \geq 0$, we have

$$\sum_{n \in \mathbb{N}} t^{n\gamma} = -(t^\gamma - 1)^{-1} \leq -\gamma^{-1} \log^{-1}(t) \exp(-\log(t)\gamma) \leq t^{-\bar{\gamma}} \log^{-1}(1/t)\gamma^{-1} ,$$

and

$$\sum_{n \in \mathbb{N}} nt^{n\gamma} = t^\gamma (t^\gamma - 1)^{-2} \leq t^\gamma \{\gamma^{-1} \log^{-1}(t) \exp(-\log(t)\gamma)\}^2 \leq t^{-\bar{\gamma}} \log^{-2}(1/t)\gamma^{-2} ,$$

which completes the proof. □

Lemma 10. *For any probability measures μ, ν on $\mathcal{B}(\mathbb{R}^d)$, measurable function $V : \mathbb{R}^d \to [1, +\infty)$ such that $\mu(V) + \nu(V) < +\infty$ and $a \in (0,1)$, we have*

$$\|\mu - \nu\|_{V^a} \leq 2\|\mu - \nu\|_V^a .$$

Proof. Let $a \in (0,1]$. The statement is trivial if $\mu = \nu$. We just need to consider the case where $\mu \neq \nu$. Define $\xi = |\mu - \nu| / (|\mu - \nu| (\mathbb{R}^d))$. Using [19, Definition D.3.1] we get that

$$\begin{aligned}
\|\mu - \nu\|_{V^a} &= (1/2)\xi(V^a) \times |\mu - \nu| (\mathbb{R}^d) \\
&\leq (1/2)\xi(V)^a \times |\mu - \nu| (\mathbb{R}^d) \\
&\leq 2^{a-1}\|\mu - \nu\|_V^a \times [|\mu - \nu| (\mathbb{R}^d)]^{1-a} ,
\end{aligned}$$

which concludes the proof using that $a \leq 1$. □

Jensen's inequality implies that **H**1-(i) holds for $V \leftarrow V^a$ with $a \in (0,1]$ since $A_1 \geq 1$. Lemma 10 implies that **H**1-(ii) holds replacing V by V^a, ρ by ρ^a and A_2 by $2A_2$. Similarly **H**1-(iii) holds replacing V by V^a and $\mathbf{\Psi}(\gamma)$ by $2\mathbf{\Psi}(\gamma)$.

C.1 Proof of Theorem 1

Consider $(\eta_n)_{n \in \mathbb{N}}$ defined for any $n \in \mathbb{N}$ by

$$\eta_n = m_n^{-1} \sum_{k=1}^{m_n} \{H_{\theta_n}(X_k^n) - \pi_{\theta_n}(H_{\theta_n})\} . \tag{25}$$

The proof of Theorem 1 relies on the two following lemmas. We consider the following decomposition for any $n \in \mathbb{N}$, $\eta_n = \eta_n^{(1)} + \eta_n^{(2)}$, where

$$\eta_n^{(1)} = \mathbb{E}\left[\eta_n | \mathcal{F}_{n-1}\right] , \qquad \eta_n^{(2)} = \eta_n - \mathbb{E}\left[\eta_n | \mathcal{F}_{n-1}\right] . \tag{26}$$

We now give upper bounds on $\mathbb{E}[\|\eta_n^{(1)}\|]$, $\mathbb{E}[\|\eta_n^{(1)}\|^2]$ and $\mathbb{E}[\|\eta_n^{(2)}\|^2]$.

Lemma 11. *Assume* **A**1*,* **A**2*,* **A**3*,* **H**1 *and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. Then we have for any $n \in \mathbb{N}$*

$$\begin{aligned}
\mathbb{E}\left[\|\eta_n^{(1)}\|\right] &\leq B_1 \mathbb{E}\left[V^{1/2}(X_0^0)\right] / (m_n\gamma_n) + \mathbf{\Psi}(\gamma_n) ; \\
\mathbb{E}\left[\|\eta_n^{(1)}\|^2\right] &\leq 2B_1^2 \mathbb{E}\left[V(X_0^0)\right] / (m_n\gamma_n)^2 + 2\mathbf{\Psi}(\gamma_n)^2 ,
\end{aligned}$$

with

$$B_1 = 2A_1A_2\rho^{-\bar{\gamma}} / \log(1/\rho) .$$

Proof. Using the definition of $(\mathcal{F}_n)_{n\in\mathbb{N}}$, see (15), the Markov property, **H**1-(ii)-(iii), Lemma 10, Jensen's inequality and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$, we have for any $n \in \mathbb{N}^*$

$$
\begin{aligned}
\|\mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right]\| &\leq m_n^{-1}\sum_{k=1}^{m_n}\left\|\mathrm{K}^k_{\gamma_n,\theta_n}H_{\theta_n}(X_0^n) - \pi_{\theta_n}\left(H_{\theta_n}\right)\right\| \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left\|\left|\delta_{X_0^n}\mathrm{K}^k_{\gamma_n,\theta_n} - \pi_{\theta_n}\right|\left(H_{\theta_n}\right)\right\| \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left|\delta_{X_0^n}\mathrm{K}^k_{\gamma_n,\theta_n} - \pi_{\theta_n}\right|\left(\|H_{\theta_n}\|\right) \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left\{\|\delta_{X_0^n}\mathrm{K}^k_{\gamma_n,\theta_n} - \pi_{\gamma_n,\theta_n}\|_{V^{1/2}}\right\} + \|\pi_{\gamma_n,\theta_n} - \pi_{\theta_n}\|_{V^{1/2}} \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left\{2A_2\rho^{k\gamma_n}V^{1/2}(X^n_{m_n}) + \boldsymbol{\Psi}(\gamma_n)\right\} \\
&\leq \frac{2A_2\rho^{-\bar{\gamma}}V^{1/2}(X^n_{m_n})}{\log(1/\rho)\gamma_n m_n} + \boldsymbol{\Psi}(\gamma_n)\ ,
\end{aligned}
$$

where for the last inequality we have used Lemma 9. In a similar manner, we have

$$
\left\|\mathbb{E}\left[\eta_0\middle|X_0^0\right]\right\| \leq \frac{2A_2\rho^{-\bar{\gamma}}V^{1/2}(X_0^0)}{\log(1/\rho)\gamma_0 m_0} + \boldsymbol{\Psi}(\gamma_0)\ .
$$

We conclude using **H**1-(i) and that $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$. $\square$

Lemma 12. *Assume* **A**1*,* **A**2*,* **A**3*,* **H**1 *and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. Then we have for any $n \in \mathbb{N}$*

$$
\mathbb{E}\left[\|\eta_n^{(2)}\|^2\right] \leq B_2 m_n^{-2}\gamma_n^{-1}\left(m_n + \gamma_n^{-1}\mathbb{E}\left[V(X_0^0)\right]\right)\ ,
$$

with $B_2 = 2(1+\bar{\gamma})^2\max(B_{2,1}, B_{2,2})$ and

$$
\begin{aligned}
B_{2,1} &= 24A_2^2(1-\rho^{1/2})^{-2}A_3\ , \\
B_{2,2} &= 4A_1\left[1 + 6A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1} + 2\right\} + A_2^2\log^{-2}(1/\rho) + A_3^2\right]\ .
\end{aligned}
$$

Proof. Let $n \in \mathbb{N}^*$. We have using the Cauchy-Schwarz inequality

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\left\{H_{\theta_n}(X_k^n) - \mathbb{E}\left[H_{\theta_n}(X_k^n)|\mathcal{F}_{n-1}\right]\right\}\right\|^2\right] \\
&\quad\leq 2\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\left\{H_{\theta_n}(X_k^n) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\}\right\|^2\right] \\
&\quad+ 2\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\left\{\mathbb{E}\left[H_{\theta_n}(X_k^n)|\mathcal{F}_{n-1}\right] - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\}\right\|^2\right]
\end{aligned}
\tag{27}
$$

Using the Markov property, **H**1-(i)-(ii), Lemma 10, Lemma 9 and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$ we obtain that

$$
\begin{aligned}
\mathbb{E}&\left[\left\|\sum_{k=1}^{m_n}\{\mathbb{E}\left[H_{\theta_n}(X_k^n)|\mathcal{F}_{n-1}\right] - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left|\sum_{k=1}^{m_n}\mathbb{E}\left[\|\delta_{X_0^n}\mathrm{R}_{\gamma_n,\theta_n} - \pi_{\gamma_n,\theta_n}\|_{V^{1/2}}\middle|\mathcal{F}_{n-1}\right]\right|^2\right] \\
&\leq 4A_2^2\mathbb{E}\left[\left|\mathbb{E}\left[V^{1/2}(X_0^n)\middle|\mathcal{F}_{n-1}\right]\sum_{k=1}^{m_n}\rho^{k\gamma_n/2}\right|^2\right] \\
&\leq 4A_1A_2^2\gamma_n^{-2}\rho^{-2\bar{\gamma}}\log^{-2}(1/\rho)\mathbb{E}\left[V(X_0^0)\right] \ . \qquad (28)
\end{aligned}
$$

We now give an upper-bound on the first term in the right-hand side of (27). Consider for any $n \in \mathbb{N}$ the Euclidean division of m_n by $\lceil 1/\gamma_n \rceil$ there exist $q_n \in \mathbb{N}$ and $r_n \in \{0, \ldots, \lceil 1/\gamma_n \rceil - 1\}$ such that $m_n = q_n \lceil 1/\gamma_n \rceil + r_n$. Therefore using the Cauchy-Schwarz inequality we can derive the following decomposition

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}H_{\theta_n}(X_k^n) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] &\leq 2\mathbb{E}\left[\left\|\sum_{j=1}^{r_n}H_{\theta_n}(X_{j+q_n\lceil 1/\gamma_n\rceil}^n) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|\sum_{j=1}^{\lceil 1/\gamma_n\rceil}\sum_{k=0}^{q_n-1}H_{\theta_n}(X_{j+k\lceil 1/\gamma_n\rceil}^n) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\sum_{j=1}^{r_n}H_{\theta_n}(\bar{X}_{q_n}^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\quad + 2\lceil 1/\gamma_n\rceil\sum_{j=1}^{\lceil 1/\gamma_n\rceil}\mathbb{E}\left[\left\|\sum_{k=0}^{q_n-1}H_{\theta_n}(\bar{X}_k^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \qquad (29)
\end{aligned}
$$

Setting for any $j \in \{1, \ldots, \lceil 1/\gamma_n \rceil\}$ and $k \in \{0, \ldots, q_n - 1\}$, $\bar{X}_k^{j,n} = X_{j+k\lceil 1/\gamma_n\rceil}^n$. We now bound the two terms in the right-hand side. First, using the Cauchy-Schwarz inequality and **H**1-(i)-(iii), the fact that $r_n \leq \lceil 1/\gamma_n \rceil$ and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$ we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^{r_n}H_{\theta_n}(\bar{X}_{q_n}^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] &\leq r_n\sum_{j=1}^{r_n}\mathbb{E}\left[\left\|H_{\theta_n}(\bar{X}_{q_n}^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\leq \lceil 1/\gamma_n\rceil^2\left(2A_1\mathbb{E}\left[V(X_0^0)\right] + 2A_3^2\right) \ . \qquad (30)
\end{aligned}
$$

Now consider the solution of the *Poisson equation* [39, Section 17.4.1] associated with $\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}$,

$x \mapsto \hat{H}_{\gamma_n,\theta_n}(x)$ defined for any $x \in \mathbb{R}^d$ by

$$\hat{H}_{\gamma_n,\theta_n}(x) = \sum_{\ell \in \mathbb{N}} \left(\mathrm{K}_{\gamma_n,\theta_n}^{\ell\lceil 1/\gamma_n \rceil} H_{\theta_n}(x) - \pi_{\gamma_n,\theta_n}(H_{\theta_n}) \right) .$$

Note that by **H**1-(ii), Lemma 10 and since for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$, we have that for any $x \in \mathbb{R}^d$

$$\left\| \hat{H}_{\gamma_n,\theta_n}(x) \right\| \leq 2A_2(1-\rho^{1/2})^{-1} V^{1/2}(x) , \tag{31}$$

and in addition for any $x \in \mathbb{R}^d$

$$\hat{H}_{\gamma_n,\theta_n}(x) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(x) = H_{\theta_n}(x) - \pi_{\gamma_n,\theta_n}(H_{\theta_n}) .$$

Therefore, we have for any $j \in \{1, \ldots, \lceil 1/\gamma_n \rceil\}$

$$\begin{aligned} \sum_{k=0}^{q_n-1} \left(H_{\theta_n}(\bar{X}_k^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n}) \right) &= \sum_{k=0}^{q_n-1} \left(\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n}) \right) \\ &= \sum_{k=0}^{q_n-2} \left(\hat{H}_{\gamma_n,\theta_n}(\bar{X}_{k+1}^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n}) \right) \\ &\quad + \hat{H}_{\gamma_n,\theta_n}(\bar{X}_0^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_{q_n-1}^{j,n}) . \end{aligned} \tag{32}$$

Combining the Cauchy-Schwarz inequality and (32) we obtain that

$$\mathbb{E}\left[\left\| \sum_{k=0}^{q_n-1} H_{\theta_n}(\bar{X}_k^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n}) \right\|^2 \right] \leq 3(\mathrm{C}_1 + \mathrm{C}_2) , \tag{33}$$

with

$$\begin{aligned} \mathrm{C}_1 &= \mathbb{E}\left[\left\| \hat{H}_{\gamma_n,\theta_n}(\bar{X}_0^{j,n}) \right\|^2 + \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \left\| \hat{H}_{\gamma_n,\theta_n}(\bar{X}_{q_n-1}^{j,n}) \right\|^2 \right] ; \\ \mathrm{C}_2 &= \mathbb{E}\left[\left\| \sum_{k=0}^{q_n-2} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_{k+1}^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n}) \right\|^2 \right] . \end{aligned}$$

First, using (31) and **H**1-(i) we get that

$$\begin{aligned} \mathrm{C}_1 &\leq 4A_2^2(1-\rho^{1/2})^{-2} \left\{ \mathbb{E}\left[V(X_j^n) \right] + \mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n} V(X_{q_n+j-1}^n) \right] \right\} \\ &\leq 8A_1A_2^2(1-\rho^{1/2})^{-2} \mathbb{E}\left[V(X_0^0) \right] . \end{aligned} \tag{34}$$

We now give an upper-bound on C_2. For any $j \in \{1, \ldots, r_n\}$ let $(\mathcal{G}_{j,k})_{k \in \{0, q_n-2\}}$ generated by $\mathcal{F}_{n-1}$ and the sequence of random variables $X_0^n, \ldots, X_{k\lceil 1/\gamma_n \rceil + j}^n$. Using the Markov property we have for any $k \in \{0, \ldots, q_n-2\}$ and $j \in \{1, \ldots, r_n\}$

$$\mathbb{E}\left[\hat{H}_{\gamma_n,\theta_n}(X_{k+1}^{j,n}) \middle| \mathcal{G}_{j,k} \right] = \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(X_k^{j,n}) .$$

Therefore, for any $j \in \{1, \ldots, r_n\}$, $\hat{H}_{\gamma_n,\theta_n}(X^{j,n}_{k+1}) - \mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \hat{H}_{\gamma_n,\theta_n}(X^{j,n}_k)$ is a martingale increment with respect to $(\mathcal{G}_{j,k})_{k\in\{0,q_n-2\}}$, Combining this result with the Markov property implies that for any $k \in \{0, \ldots, q_n - 2\}$ and $j \in \{1, \ldots, r_n\}$,

$$
\begin{aligned}
\mathrm{C}_2 &= \sum_{k=0}^{q_n-2} \mathbb{E}\left[\mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k) - \mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k)\right\|^2\right] \\
&= \sum_{k=0}^{q_n-2} \mathbb{E}\left[\mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k)\right\|^2 - \left\|\mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k)\right\|^2\right] . \qquad (35)
\end{aligned}
$$

Define for any $x \in \mathbb{R}^d$, $g_n(x) = \|\hat{H}_{\gamma_n,\theta_n}(x)\|^2$. Using (35), **H**1-(ii)-(iii) and (31) we obtain that

$$
\begin{aligned}
\mathrm{C}_2 &= \sum_{k=0}^{q_n-2} \mathbb{E}\left[\mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k)\right\|^2 - \left\|\mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k)\right\|^2\right] \\
&\leq \sum_{k=0}^{q_n-2} \mathbb{E}\left[\mathrm{K}^{\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} \left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}^{j,n}_k)\right\|^2\right] \\
&\leq \mathbb{E}\left[\sum_{k=0}^{q_n-2} \mathbb{E}\left[\mathrm{K}^{(k+1)\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} g_n(\bar{X}^{j,n}_0) - \pi_{\gamma_n,\theta_n}(g_n) \middle| \mathcal{G}_{j,0}\right]\right] + \sum_{k=0}^{q_n-2} \pi_{\gamma_n,\theta_n}(g_n) \\
&\leq \frac{4A_2^2}{(1-\rho^{1/2})^2} \left\{\sum_{k=0}^{q_n-2} \mathbb{E}\left[\mathbb{E}\left[\|\delta_{X^n_j} \mathrm{K}^{(k+1)\lceil 1/\gamma_n \rceil}_{\gamma_n,\theta_n} - \pi_{\gamma_n,\theta_n}\|_V \middle| \mathcal{G}_{j,0}\right]\right] + \sum_{k=0}^{q_n-2} \pi_{\gamma_n,\theta_n}(V)\right\} \\
&\leq 4A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1}\mathbb{E}\left[V(X^n_j)\right] + q_n A_3\right\} \\
&\leq 4A_2^2(1-\rho^{1/2})^{-2}\left\{A_1 A_2(1-\rho)^{-1}\mathbb{E}\left[V(X^0_0)\right] + q_n A_3\right\} . \qquad (36)
\end{aligned}
$$

Therefore, using (34) and (36) in (33) we obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=0}^{q_n-1} H_{\theta_n}(\bar{X}^{j,n}_k) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\qquad \leq 12A_2^2(1-\rho^{1/2})^{-2}\left[\left\{A_1A_2(1-\rho)^{-1}\mathbb{E}\left[V(X^0_0)\right] + q_nA_3\right\} + 2\mathbb{E}\left[V(X^0_0)\right]\right] . \qquad (37)
\end{aligned}
$$

As a consequence, using (30) and (37) in (29) we get that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=1}^{m_n} H_{\theta_n}(X^n_k) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \leq 4\lceil 1/\gamma_n \rceil^2 \left(A_1\mathbb{E}\left[V(X^0_0)\right] + A_3^2\right) \\
&\quad + 24\lceil 1/\gamma_n \rceil^2 A_2^2(1-\rho^{1/2})^{-2}\left\{A_1\mathbb{E}\left[V(X^0_0)\right](A_2(1-\rho)^{-1}+2) + q_nA_3\right\} \\
&\leq \left[\gamma_n^{-2}\left(A_1\mathbb{E}\left[V(X^0_0)\right]\left[24A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1}+2\right\}+4\right] + 4A_3^2\right)\right. \\
&\left. +24A_2^2(1-\rho^{1/2})^{-2}A_3m_n/\gamma_n\right](1+\bar{\gamma})^2 \qquad (38)
\end{aligned}
$$

Combining (28) and (38) in (27) we obtain that

$$\mathbb{E}\left[\left\|\sum_{k=1}^{m_n} H_{\theta_n}(X_k^n) - \mathbb{E}\left[H_{\theta_n}(X_k^n)\right]\right\|^2\right] \leq 8\gamma_n^{-2} A_1 A_2^2 \rho^{-2\bar{\gamma}} \log^{-2}(1/\rho) \mathbb{E}\left[V(X_0^0)\right]$$
$$+ 2\left[\gamma_n^{-2}\left(A_1 \mathbb{E}\left[V(X_0^0)\right]\left[24A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1}+2\right\}+4\right]+4A_3^2\right)\right.$$
$$\left.+24A_2^2(1-\rho^{1/2})^{-2}A_3 m_n/\gamma_n\right](1+\bar{\gamma})^2$$
$$\leq 2(1+\bar{\gamma})^2\left(A_1\mathbb{E}\left[V(X_0^0)\right]\left[24A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1}+2\right\}\right.\right.$$
$$\left.\left.+4\left\{1+A_2^2\log^{-2}(1/\rho)\right\}\right]+4A_3^2\right)\gamma_n^{-2}+48A_2^2(1-\rho^{1/2})^{-2}A_3(1+\bar{\gamma})^2(m_n/\gamma_n)\ ,$$

which concludes the proof for $n \neq 0$. The same inequality holds in the case where $n = 0$. □

We now turn to the proof of Theorem 1.

Proof of Theorem 1. The proof is an application of [2, Theorem 2, Theorem 3].

(a) To apply [2, Theorem 2], it is enough to show that the following series converge almost surely

$$\sum_{n=0}^{+\infty} \delta_{n+1}\langle \Pi_\Theta(\theta_n - \delta_{n+1}\nabla f(\theta_n)), \eta_n^{(i)}\rangle\ ,\quad \sum_{n=0}^{+\infty}\delta_{n+1}\eta_n^{(i)}\ ,\quad \sum_{n=0}^{+\infty}\delta_{n+1}^2\|\eta_n^{(i)}\|^2\ .$$

where $i \in \{1,2\}$ and the sequences $(\eta_n^{(1)})_{n\in\mathbb{N}}$ and $(\eta_n^{(2)})_{n\in\mathbb{N}}$ are given in (27).

In the case where $i = 1$, since $(\Pi_\Theta(\theta_n - \delta_{n+1}\nabla f(\theta_n)))_{n\in\mathbb{N}}$ is bounded, we are reduced to proving that almost surely $\sum_{n=0}^{+\infty}\delta_{n+1}\|\eta_n^{(1)}\| < +\infty$. Using (16), Lemma 11 and Fubini-Tonelli's theorem we obtain that

$$\mathbb{E}\left[\sum_{n\in\mathbb{N}}\delta_{n+1}\|\eta_n^{(1)}\|\right] = \sum_{n\in\mathbb{N}}\delta_{n+1}\mathbb{E}\left[\|\eta_n^{(1)}\|\right] < +\infty\ . \tag{39}$$

We consider the case where $i = 2$. Let $(S_n)_{n\in\mathbb{N}}$ and $(T_n)_{n\in\mathbb{N}}$ be defined for any $n \in \mathbb{N}$ by $S_n = \sum_{k=0}^n \delta_{k+1}\langle\Pi_\Theta(\theta_k - \delta_{k+1}\nabla f(\theta_k)), \eta_k^{(2)}\rangle$ and $T_n = \sum_{k=0}^n \delta_{k+1}\eta_n^{(2)}$ are $(\mathcal{F}_n)_{n\in\mathbb{N}}$-martingale by definition of $(\eta_n^{(2)})_{n\in\mathbb{N}}$ in (27) and $(\mathcal{F}_n)_{n\in\mathbb{N}}$ in (15). Therefore, using [56, Section 12.5], the Cauchy-Schwarz inequality and that the sequence $(\Pi_\Theta(\theta_n - \delta_{n+1}\nabla f(\theta_n)))_{n\in\mathbb{N}}$ is bounded, it suffices to show that $\sum_{n=0}^{+\infty}\delta_{n+1}^2\mathbb{E}[\|\eta_n^{(2)}\|^2] < +\infty$. Using Lemma 12 we get that

$$\sum_{n=0}^{+\infty}\delta_{n+1}^2\mathbb{E}[\|\eta_n^{(2)}\|^2] \leq B_2\left(\sum_{n=0}^{+\infty}\delta_{n+1}^2/(m_n\gamma_n) + \mathbb{E}\left[V(X_0^0)\right]\sum_{n=0}^{+\infty}\delta_{n+1}^2/(m_n\gamma_n)^2\right)\ .$$

Combining this result and (39) implies the stated convergence applying [2, Theorem 2].

(b) Applying [2, Theorem 3], the Cauchy-Schwarz inequality and using **A**1 we obtain that almost surely for any $n \in \mathbb{N}$

$$
\begin{aligned}
&\sum_{k=1}^{n} \delta_k \left\{ f(\theta_k) - \min_{\Theta} f \right\} \\
&\quad \le \frac{\|\theta_0 - \theta^\star\|^2}{2} - \sum_{k=0}^{n-1} \delta_{k+1} \langle \Pi_\Theta(\theta_k - \delta_{k+1} \nabla f(\theta_k)) - \theta^\star, \eta_k \rangle + \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k\|^2 \\
&\quad \le 2M_\Theta^2 - \sum_{i=1}^{2} \sum_{k=0}^{n-1} \delta_{k+1} \langle \Pi_\Theta(\theta_k - \delta_{k+1} \nabla f(\theta_k)) - \theta^\star, \eta_k^{(i)} \rangle + 2 \sum_{\hat{\imath}=1}^{2} \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k^{(i)}\|^2 \; . \qquad (40)
\end{aligned}
$$

which implies by the proof of (a) that $\sup_{n \in \mathbb{N}} [\sum_{k=1}^{n} \delta_k \left\{ f(\theta_k) - \min_\Theta f \right\}] < +\infty$ almost surely. The proof is then completed upon dividing (40) by $\sum_{k=1}^{n} \delta_k$.

□

C.2 Proof of Theorem 2

Proof. Taking the expectation in (40) and using that $\eta_n^{(2)}$ is a martingale increment with respect to $(\mathcal{F}_n)_{n \in \mathbb{N}}$, we get that for every $n \in \mathbb{N}$

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{n} \delta_k \left\{ f(\theta_k) - \min_{\Theta} f \right\} \right] \\
&\le \mathbb{E}\left[2M_\Theta^2 - \sum_{k=0}^{n-1} \delta_{k+1} \langle \Pi_\Theta(\theta_k - \delta_{k+1} \nabla f(\theta_k)) - \theta^\star, \eta_k \rangle + \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k\|^2 \right] \\
&\le 2M_\Theta^2 + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1} \mathbb{E}\left[\left\| \eta_k^{(1)} \right\| \right] + 2 \sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbb{E}\left[\left\| \eta_k^{(1)} \right\|^2 \right] + 2 \sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbb{E}\left[\left\| \eta_k^{(2)} \right\|^2 \right]
\end{aligned}
$$

Combining this result, Lemma 11 and Lemma 12 completes the proof. □

C.3 Proof of Theorem 3

We now introduce some tools needed for the proof. By **A**4 and **H**1-(i)-(ii), for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, there exists a function $\hat{H}_{\gamma,\theta} : \mathbb{R}^d \to \mathbb{R}^{d_\theta}$ solution of the *Poisson equation*,

$$
(\mathrm{Id} - \mathrm{K}_{\gamma,\theta}) \hat{H}_{\gamma,\theta} = H_\theta - \pi_{\gamma,\theta}(H_\theta) \; , \qquad (41)
$$

defined for any $x \in \mathbb{R}^d$ by

$$
\hat{H}_{\gamma,\theta}(x) = \sum_{j \in \mathbb{N}} \{ \mathrm{K}_{\gamma,\theta}^j H_\theta(x) - \pi_{\gamma,\theta}(H_\theta) \} \; . \qquad (42)
$$

Note that using **H**1-(ii) and Lemma 10 we have for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$

$$
\left\| \hat{H}_\theta(x) \right\| \le C_{\hat{H}} \gamma^{-1} V^{1/4}(x) \; , \qquad C_{\hat{H}} = 8A_2 \log^{-1}(1/\rho) \rho^{-\bar{\gamma}/4} \; . \qquad (43)
$$

Define for any $n \in \mathbb{N}$

$$\tilde{\eta}_n = H_{\theta_n}(\tilde{X}_{n+1}) - \pi_{\tilde{\theta}_n}(H_{\tilde{\theta}_n}) . \tag{44}$$

Using (41) an alternative expression of $(\tilde{\eta}_n)_{n\in\mathbb{N}}$ is given for any $n \in \mathbb{N}$ by

$$\begin{aligned}\tilde{\eta}_n &= \hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) + \pi_{\gamma_n,\tilde{\theta}_n}(H_{\tilde{\theta}_n}) - \pi_{\tilde{\theta}_n}(H_{\tilde{\theta}_n}) \\ &= \tilde{\eta}_n^{(a)} + \tilde{\eta}_n^{(b)} + \tilde{\eta}_n^{(c)} + \tilde{\eta}_n^{(d)} ,\end{aligned}$$

where

$$\begin{aligned}\tilde{\eta}_n^{(a)} &= \hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n) , \\ \tilde{\eta}_n^{(b)} &= \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n) - \mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1}) , \\ \tilde{\eta}_n^{(c)} &= \mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) , \\ \tilde{\eta}_n^{(d)} &= \pi_{\gamma_n,\tilde{\theta}_n}(H_{\tilde{\theta}_n}) - \pi_{\tilde{\theta}_n}(H_{\tilde{\theta}_n}) .\end{aligned} \tag{45}$$

To establish Theorem 3 we need to get estimates on moments of $\left\|\tilde{\eta}_n^{(i)}\right\|$ for $i \in \{a,b,c,d\}$. It is the matter of the following technical results.

Lemma 13. *Assume* **A**1*,* **A**2*,* **A**3*,* **H**1 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\|H_\theta(x)\| \le V^{1/4}(x)$*. Then we have for any* $n \in \mathbb{N}$*,* $\mathbb{E}\left[\|\tilde{\eta}_n\|^2\right] \le C_1$*, with*

$$C_1 = 2A_1\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] + 2\sup_{\theta\in\Theta}\|\nabla f(\theta)\|^2 .$$

Proof. Using (44), that $\|x+y\|^2 \le 2(\|x\|^2 + \|y\|^2)$ for any $x, y \in \mathbb{R}^d$, **A**1, **A**2, **A**3 and **H**1-(i) and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/2}(x)$, we get for any $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\eta}_k\|^2\right] &\le 2\mathbb{E}[\|H_{\tilde{\theta}_k}(\tilde{X}_{k+1})\|^2] + 2\left[\pi_{\tilde{\theta}_k}(\|H_{\tilde{\theta}_k}\|)\right]^2 \\ &\le 2A_1\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] + 2\sup_{\theta\in\Theta}\|\nabla f(\theta)\|^2 .\end{aligned}$$

□

Lemma 14. *Assume* **A**1*,* **A**2*,* **A**3*,* **A**4*,* **H**1*,* **H**2 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\|H_\theta(x)\| \le V^{1/4}(x)$*. Then we have for any* $n \in \mathbb{N}$*,* $\mathbb{E}\left[\left\|\tilde{\eta}_n^{(a)}\right\|^2\right] \le \tilde{C}_1\gamma_n^{-2}$*, with*

$$\tilde{C_1} = A_1 C_{\hat{H}}^2\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] .$$

Proof. By (45), using (43) and **H**1-(i) we get that for any $n \in \mathbb{N}^*$

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\left\|\tilde{\eta}_n^{(a)}\right\|^2\middle|\mathcal{F}_n\right]\right] \\ &\le \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1})\right\|^2\middle|\mathcal{F}_n\right]\right] - \mathbb{E}\left[\left\|\mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n)\right\|^2\right] \\ &\le A_1 C_{\hat{H}}^2\gamma_n^{-2}\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] ,\end{aligned}$$

which concludes the proof. □

Lemma 15. *Assume* **A**1*,* **A**2*,* **A**3*,* **H**1 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\|H_\theta(x)\| \le V^{1/4}(x)$*. Then the following statements hold.*

(a) There exists $C_3 \ge 0$ *such that for any* $n \in \mathbb{N}$ *and* $\theta \in \Theta$

$$\mathbb{E}\left[\left\|\sum_{k=0}^{n} \delta_{k+1}\langle a_{k+1}, \tilde{\eta}_k^{(b)}\rangle\right\|\right] \le C_3 \left[\sum_{k=0}^{n} |\delta_{k+1} - \delta_k|\, \gamma_k^{-1} + \sum_{k=0}^{n} \delta_{k+1}^2 \gamma_k^{-1} + (\delta_{n+1}/\gamma_{n+1} - \delta_1/\gamma_1)\right] .$$

with $a_{k+1} = \Pi_\Theta\left[\tilde{\theta}_k - \delta_{k+1}\nabla f(\tilde{\theta}_k)\right] - \theta^\star$*,* $\theta^\star \in \arg\min_\Theta f$ *and*

$$C_3 = A_1 C_{\hat{H}}(4M_\Theta + \sup_{\theta \in \Theta}\|\nabla f(\theta)\| + 1 + \delta_1 L_f)\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right] .$$

(b) If (22) *holds then* $\sum_{k=0}^{n} \delta_{k+1}\langle a_{k+1}, \tilde{\eta}_k^{(b)}\rangle$ *converges almost surely.*

Proof. By (45) we have for any $n \in \mathbb{N}$ and $\theta \in \Theta$

$$\begin{aligned}
&\sum_{k=0}^{n} \delta_{k+1}\langle a_{k+1}, \tilde{\eta}_k^{(b)}\rangle \\
&= \sum_{k=0}^{n} \langle \delta_{k+1} a_{k+1}, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k) - \mathrm{K}_{\gamma_{k+1},\tilde{\theta}_{k+1}}\hat{H}_{\gamma_{k+1},\tilde{\theta}_{k+1}}(\tilde{X}_{k+1})\rangle \\
&= \sum_{k=1}^{n} \langle \delta_{k+1} a_{k+1} - \delta_k a_k, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k)\rangle \\
&- \langle \delta_{n+1} a_{n+1}, \mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1})\rangle \\
&+ \langle \delta_1 a_1, \mathrm{K}_{\gamma_0,\tilde{\theta}_0}\hat{H}_{\gamma_0,\tilde{\theta}_0}(\tilde{X}_0)\rangle ,
\end{aligned} \tag{46}$$

In addition, we have for any $n \in \mathbb{N}$, $\theta \in \Theta$ using **A**1, **A**2, that Π_Θ is non-expansive, (20), **H**1-(i) and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$

$$\begin{aligned}
\|\delta_{n+1} a_{n+1} - \delta_n a_n\| &\le 2M_\Theta |\delta_{n+1} - \delta_n| + \delta_{n+1}\|a_{n+1} - a_n\| \\
&\le 2M_\Theta |\delta_{n+1} - \delta_n| + (1 + \delta_n L_f)\|\theta_{n+1} - \theta_n\| + |\delta_{n+1} - \delta_n|\,\|\nabla f(\theta_{n+1})\| \\
&\le (2M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\|)\,|\delta_{n+1} - \delta_n| + \delta_{n+1}^2(1 + \delta_{n+1} L_f) V^{1/4}(\tilde{X}_{n+1}) .
\end{aligned} \tag{47}$$

(a) Combining (46), (47), (43), the Cauchy-Schwarz inequality and **H**1-(i) we get that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=0}^{n} \delta_{k+1}\langle a_k, \tilde{\eta}_k^{(b)}\rangle\right\|\right] &\le (2M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\|) A_1 C_{\hat{H}} \mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right] \sum_{k=0}^{n} |\delta_{k+1} - \delta_k|\,\gamma_k^{-1} \\
&+ A_1 C_{\hat{H}}(1 + \delta_1 L_f)\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right] \sum_{k=0}^{n} \delta_{k+1}^2 \gamma_k^{-1} \\
&+ 2A_1 M_\Theta C_{\hat{H}} \mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right] \{\delta_{n+1}/\gamma_{n+1} + \delta_1/\gamma_1\} ,
\end{aligned}$$

which concludes the proof of Lemma 15-(a).

(b) Assume now (22). We show that almost surely the first term in (46) is absolutely convergence and the second term converges to 0.

Using (47), (43), the Cauchy-Schwarz inequality and (22) we get that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{+\infty}\left|\langle \delta_{k+1}a_{k+1}-\delta_k a_k, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k)\rangle\right|\right] \\
&\leq (2M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\|)A_1 C_{\hat{H}}\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\sum_{k=0}^{+\infty}|\delta_{k+1}-\delta_k|\,\gamma_k^{-1} \\
&\qquad + A_1 C_{\hat{H}}(1+\delta_1 L_f)\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\sum_{k=0}^{+\infty}\delta_{k+1}^2 < +\infty \;,
\end{aligned}$$

which implies that $(\langle \delta_{k+1}a_{k+1}-\delta_k a_k, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k)\rangle)_{k\in\mathbb{N}}$ is absolutely convergent almost surely.

We have that $\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\|\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1})\|$ is upper-bounded using (43) by $\gamma_{n+1}^{-1}C_{\hat{H}}\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}V^{1/4}(\tilde{X}_{n+1})$. It follows that we have for any $\theta\in\Theta$, $\varepsilon>0$, using the Markov inequality, the Cauchy-Schwarz inequality, (43) and (22)

$$\begin{aligned}
&\sum_{n\in\mathbb{N}}\mathbb{P}\left(\|a_{n+1}\|\,\delta_{n+1}\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\|\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1})\|\geq\varepsilon\right) \\
&\qquad\leq \sum_{n\in\mathbb{N}}\mathbb{P}\left(2C_{\hat{H}}M_\Theta\,\delta_{n+1}\,\gamma_{n+1}^{-1}\,\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}V^{1/4}(\tilde{X}_{n+1})\geq\varepsilon\right) \\
&\qquad\leq 4\varepsilon^{-2}M_\Theta^2 C_{\hat{H}}^2 A_1\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right]\sum_{n\in\mathbb{N}}\delta_n^2\gamma_n^{-2} < +\infty\;,
\end{aligned}$$

Using the Borel-Cantelli lemma, we get $\lim_{n\to+\infty}\langle\delta_n a_n \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n)\rangle = 0$ almost surely. This completes the proof of convergence of the series $\sum_{k\in\mathbb{N}}\delta_{k+1}\langle a_{k+1},\tilde{\eta}_k^{(b)}\rangle$ for any $\theta\in\Theta$.

□

Lemma 16. *Assume* **A**1*,* **A**2*,* **A**3*,* **A**4*,* **H**1*,* **H**2 *and that for any $\theta\in\Theta$ and $x\in\mathbb{R}^d$, $\|H_\theta(x)\|\leq V^{1/4}(x)$. Then we have for any $n\in\mathbb{N}$*

$$\mathbb{E}\left[\left\|\tilde{\eta}_n^{(c)}\right\|\right] \leq C_2\gamma_{n+1}^{-1}\left[\gamma_{n+1}^{-1}\left\{\boldsymbol{\Lambda}_1(\gamma_n,\gamma_{n+1})+\boldsymbol{\Lambda}_2(\gamma_n,\gamma_{n+1})\delta_{n+1}\right\}+\delta_{n+1}\right]\;,$$

with

$$C_2 = 4A_1A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\max\left[L_H, C_{c,1}+2A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\right]\;, \tag{48}$$

where $C_{c,1}$ is given by

$$C_{c,1} = 4A_1A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\mathbb{E}\left[V(\tilde{X}_0)\right]\;. \tag{49}$$

Proof. We start by giving an upper-bound on $\|\pi_{\gamma_1,\theta_1}-\pi_{\gamma_2,\theta_2}\|_{V^{1/2}}$ for $\gamma_1,\gamma_2\in(0,\bar{\gamma}]$ with $\gamma_1>\gamma_2$ and, $\theta_1,\theta_2\in\Theta$. Let $g:\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ be a measurable function satisfying $\sup_{x\in\mathbb{R}^d}\{\|g(x)\|/V^{1/2}(x)\}\leq$

1. Using **H**1-(i)-(ii), **H**2, Lemma 9 and Lemma 10, we get that for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $\theta_1, \theta_2 \in \Theta$ and $\ell \in \mathbb{N}^*$

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathrm{K}^{\ell}_{\gamma_1,\theta_1} g(\tilde{X}_0) - \mathrm{K}^{\ell}_{\gamma_2,\theta_2} g(\tilde{X}_0)\right\|\right] \\
&= \left\| \sum_{j=0}^{\ell-1} \mathrm{K}^{j}_{\gamma_1,\theta_1} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{ \mathrm{K}^{(\ell-1-j)}_{\gamma_2,\theta_2} g(x) - \pi_{\gamma_2,\theta_2}(f) \right\} \right\| \\
&\le 2A_2 \sum_{j=0}^{\ell-1} \rho^{(\ell-1-j)\gamma_2/2} \left\| \mathrm{K}^{j}_{\gamma_1,\theta_1} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) V^{1/2}(x) \right\| \\
&\le 2A_2 \sum_{j=0}^{\ell-1} \rho^{(\ell-1-j)\gamma_2/2} \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2) \|\theta_1 - \theta_2\| \right] \sup_{k \in \mathbb{N}} \mathbb{E}\left[\mathrm{K}^{k}_{\gamma_1,\theta_1} V(\tilde{X}_0) \right] \\
&\le 4A_1 A_2 \log^{-1}(1/\rho) \rho^{-\bar{\gamma}/2} \gamma_2^{-1} \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2) \|\theta_1 - \theta_2\| \right] \mathbb{E}\left[V(\tilde{X}_0) \right] \ .
\end{aligned}
$$

Taking $\ell \to +\infty$ and using **H** 1-(ii), we obtain that for any $\theta_1, \theta_2 \in \Theta$ and $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$,

$$
\|\pi_{\gamma_1,\theta_1} - \pi_{\gamma_2,\theta_2}\|_{V^{1/2}} \le C_{c,1} \gamma_2^{-1} \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2) \|\theta_1 - \theta_2\| \right] \ , \tag{50}
$$

with $C_{c,1}$ given by(49). In what follows we give an upper bound on $\left\| \mathrm{K}_{\gamma_1,\theta_1} \hat{H}_{\gamma_1,\theta_1}(x) - \mathrm{K}_{\gamma_2,\theta_2} \hat{H}_{\gamma_2,\theta_2}(x) \right\|$ for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$ and $x \in \mathbb{R}^d$. By (42) we have for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$ and $x \in \mathbb{R}^d$,

$$
\begin{aligned}
&\left\| \mathrm{K}_{\gamma_1,\theta_1} \hat{H}_{\gamma_1,\theta_1}(x) - \mathrm{K}_{\gamma_2,\theta_2} \hat{H}_{\gamma_2,\theta_2}(x) \right\| \\
&= \left\| \sum_{\ell \in \mathbb{N}^*} \left\{ \mathrm{K}^{\ell}_{\gamma_1,\theta_1} H_{\theta_1}(x) - \pi_{\gamma_1,\theta_1}(H_{\theta_1}) \right\} - \sum_{\ell \in \mathbb{N}^*} \left\{ \mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_2}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_2}) \right\} \right\| \\
&\le \sum_{\ell \in \mathbb{N}^*} \left\| \left\{ \mathrm{K}^{\ell}_{\gamma_1,\theta_1} H_{\theta_1}(x) - \pi_{\gamma_1,\theta_1}(H_{\theta_1}) \right\} - \left\{ \mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_2}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_2}) \right\} \right\| \ .
\end{aligned}
$$

We now bound each term of the series in the right hand side. For any measurable functions g_1, g_2 with $g_i : \mathbb{R}^d \to \mathbb{R}^{d_\theta}$ and such that $\sup_{x \in \mathbb{R}^d} \|g_i(x)\| / V^{1/4}(x) < +\infty$ with $i \in \{1, 2\}$, $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^*$, it holds that

$$
\begin{aligned}
&\mathrm{K}^{\ell}_{\gamma_1,\theta_1} g_1(x) - \mathrm{K}^{\ell}_{\gamma_2,\theta_2} g_2(x) = \mathrm{K}^{\ell}_{\gamma_1,\theta_1} g_1(x) - \mathrm{K}^{\ell}_{\gamma_2,\theta_2} g_1(x) + \mathrm{K}^{\ell}_{\gamma_2,\theta_2} (g_1(x) - g_2(x)) \\
&= \sum_{j=0}^{\ell-1} \left\{ \mathrm{K}^{j}_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1} \right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{ \mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} g_1(x) - \pi_{\gamma_2,\theta_2}(g_1) \right\} \\
&\quad + \sum_{j=0}^{\ell-1} \pi_{\gamma_1,\theta_1} \left\{ \mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} g_1(x) - \mathrm{K}^{\ell-j}_{\gamma_2,\theta_2} g_1(x) \right\} + \mathrm{K}^{\ell}_{\gamma_2,\theta_2} (g_1(x) - g_2(x)) \\
&= \sum_{j=0}^{\ell-1} \left\{ \mathrm{K}^{j}_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1} \right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{ \mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} g_1(x) - \pi_{\gamma_2,\theta_2}(g_1) \right\} \\
&\quad - \pi_{\gamma_1,\theta_1} (\mathrm{K}^{\ell}_{\gamma_2,\theta_2} g_1(x) - g_1(x)) + \mathrm{K}^{\ell}_{\gamma_2,\theta_2} (g_1(x) - g_2(x)) \ .
\end{aligned}
$$

Setting $g_1 = H_{\theta_1} - \pi_{\gamma_1,\theta_1}(H_{\theta_1})$ and $g_2 = H_{\theta_2} - \pi_{\gamma_2,\theta_2}(H_{\theta_2})$, we obtain that

$$\begin{aligned}
&\mathrm{K}_{\gamma_1,\theta_1}^{\ell} g_1(x) - \mathrm{K}_{\gamma_2,\theta_2}^{\ell} g_2(x) \\
&\qquad = \sum_{j=0}^{\ell-1} \left\{\mathrm{K}_{\gamma_1,\theta_1}^{j} - \pi_{\gamma_1,\theta_1}\right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{\mathrm{K}_{\gamma_2,\theta_2}^{\ell-1-j} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})\right\} + \Xi_\ell \;, \qquad (51)
\end{aligned}$$

where

$$\begin{aligned}
\Xi_\ell &= -\pi_{\gamma_1,\theta_1}(\mathrm{K}_{\gamma_2,\theta_2}^{\ell} H_{\theta_1}(x) - H_{\theta_1}(x)) \\
&\quad + \mathrm{K}_{\gamma_2,\theta_2}^{\ell}\left[H_{\theta_1}(x) - H_{\theta_2}(x) + \pi_{\gamma_2,\theta_2}(H_{\theta_2}) - \pi_{\gamma_1,\theta_1}(H_{\theta_1})\right] \\
&\quad = -\pi_{\gamma_1,\theta_1}\mathrm{K}_{\gamma_2,\theta_2}^{\ell} H_{\theta_1}(x) + \mathrm{K}_{\gamma_2,\theta_2}^{\ell}\left[H_{\theta_1}(x) - H_{\theta_2}(x) + \pi_{\gamma_2,\theta_2}(H_{\theta_2})\right] \\
&\quad = (\pi_{\gamma_2,\theta_2} - \pi_{\gamma_1,\theta_1})(\mathrm{K}_{\gamma_2,\theta_2}^{\ell} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})) - \pi_{\gamma_2,\theta_2}(H_{\theta_1}) \\
&\qquad\qquad + \mathrm{K}_{\gamma_2,\theta_2}^{\ell}\left[H_{\theta_1}(x) - H_{\theta_2}(x) + \pi_{\gamma_2,\theta_2}(H_{\theta_2})\right] \\
&\quad = (\pi_{\gamma_2,\theta_2} - \pi_{\gamma_1,\theta_1})(\mathrm{K}_{\gamma_2,\theta_2}^{\ell} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})) \qquad (52) \\
&\qquad\qquad + \mathrm{K}_{\gamma_2,\theta_2}^{\ell}(H_{\theta_1} - H_{\theta_2})(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1} - H_{\theta_2}) \;.
\end{aligned}$$

For the first term in (51), using **H**1-(ii), **H**2, Lemma 10 and and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x)$ we obtain for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^*$

$$\begin{aligned}
&\left\| \sum_{j=0}^{\ell-1} \left\{\mathrm{K}_{\gamma_1,\theta_1}^{j} - \pi_{\gamma_1,\theta_1}\right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{\mathrm{K}_{\gamma_2,\theta_2}^{\ell-1-j} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})\right\} \right\| \\
&\qquad \leq 2A_2 \sum_{j=0}^{\ell-1} \rho^{(\ell-1-j)\gamma_1/2} \left\| \left\{\mathrm{K}_{\gamma_1,\theta_1}^{j} - \pi_{\gamma_1,\theta_1}\right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) V^{1/2}(x) \right\| \\
&\qquad \leq 4A_2^2 \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right] \sum_{j=0}^{\ell-1} \rho^{(j+(\ell-1-j))\gamma_2/2} V^{1/2}(x) \\
&\qquad \leq 4A_2^2 \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right] \ell \rho^{(\ell-1)\gamma_2/2} V^{1/2}(x) \;. \qquad (53)
\end{aligned}$$

For the first term in (52), using **H**1-(ii), Lemma 10, (50) and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x) \leq V^{1/2}(x)$, we obtain for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^*$

$$\begin{aligned}
&\left\|(\pi_{\gamma_1,\theta_1} - \pi_{\gamma_2,\theta_2})(\mathrm{K}_{\gamma_2,\theta_2}^{\ell} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1}))\right\| \\
&\qquad \leq 2A_2 \rho^{\ell\gamma_2/2} \|\pi_{\gamma_1,\theta_1} - \pi_{\gamma_2,\theta_2}\|_{V^{1/2}} \\
&\qquad \leq 2A_2 C_{c,1}\, \rho^{\ell\gamma_2/2} \gamma_2^{-1} \left\{\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right\} \;. \qquad (54)
\end{aligned}$$

For the second term in (52), using **A**4, **H**1-(ii) and Lemma 10, we obtain for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^*$

$$\left\|\mathrm{K}_{\gamma_2,\theta_2}^{\ell}(H_{\theta_1} - H_{\theta_2})(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1} - H_{\theta_2})\right\| \leq 2A_2 L_H \rho^{\ell\gamma_2/2} \|\theta_1 - \theta_2\| V^{1/2}(x) \;. \qquad (55)$$

Combining (52), (53), (54), (55) in (51) and using Lemma 9, we obtain that for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ that

$$\begin{aligned}&\left\| \mathrm{K}_{\gamma_1,\theta_1} \hat{H}_{\gamma_1,\theta_1}(x) - \mathrm{K}_{\gamma_2,\theta_2} \hat{H}_{\gamma_2,\theta_2}(x) \right\| \\ &\qquad\qquad \le C_{c,2}\, \gamma_2^{-1} \left[\gamma_2^{-1} \left\{ \mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2) \|\theta_1 - \theta_2\| \right\} + \|\theta_1 - \theta_2\| \right] V^{1/2}(x) \;,\end{aligned}$$

with

$$C_{c,2} = 4A_2 \log^{-1}(1/\rho) \rho^{-\bar{\gamma}/2} \max \left[L_H, C_{c,1} + 2A_2 \log^{-1}(1/\rho) \rho^{-\bar{\gamma}/2} \right] \;.$$

Since for any $k \in \mathbb{N}$, $\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\| \le \delta_{k+1} V^{1/2}(\tilde{X}_{k+1})$ by (20) and the fact that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/2}(x)$ and that Π_Θ is non-expansive, we get that for any $k \in \mathbb{N}$,

$$\begin{aligned}&\left\| \mathrm{K}_{\gamma_k,\tilde{\theta}_k} \hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_{k+1}) - \mathrm{K}_{\gamma_{k+1},\tilde{\theta}_{k+1}} \hat{H}_{\gamma_{k+1},\tilde{\theta}_{k+1}}(\tilde{X}_{k+1}) \right\| \\ &\qquad\qquad \le C_{c,2} \gamma_{k+1}^{-1} \left[\gamma_{k+1}^{-1} \left\{ \mathbf{\Lambda}_1(\gamma_k, \gamma_{k+1}) + \mathbf{\Lambda}_2(\gamma_k,\gamma_{k+1}) \delta_{k+1} \right\} + \delta_{k+1} \right] V(\tilde{X}_{k+1}) \;,\end{aligned}$$

which implies by (45) and using **H**1-(i) that

$$\mathbb{E}\left[\left\| \tilde{\eta}^{(c)} \right\| \right] \le C_2 \gamma_{k+1}^{-1} \left[\gamma_{k+1}^{-1} \left\{ \mathbf{\Lambda}_1(\gamma_k,\gamma_{k+1}) + \mathbf{\Lambda}_2(\gamma_k,\gamma_{k+1}) \delta_{k+1} \right\} + \delta_{k+1} \right] \;,$$

with C_2 given by (48).

□

Lemma 17. *Assume* **A**1*,* **A**2*,* **A**3*,* **H**1 *and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$. Then we have for any $n \in \mathbb{N}$*

$$\mathbb{E}\left[\left\| \tilde{\eta}_n^{(d)} \right\| \right] \le \mathbf{\Psi}(\gamma_n) \;.$$

Proof. By a straightforward application of **H**1-(iii) and by (45) along with the fact that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$ we have for any $n \in \mathbb{N}$, $\mathbb{E}\left[\left\| \tilde{\eta}_n^d \right\| \right] \le \mathbf{\Psi}(\gamma_n)$. □

We now turn to the proof of Theorem 3.

Proof of Theorem 3. (a) To apply [2, Theorem 2], it is enough to show that the following series converge almost surely

$$\sum_{n=0}^{+\infty} \delta_{n+1} \langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)) - \theta^\star, \tilde{\eta}_n^{(i)} \rangle \;, \quad \sum_{n=0}^{+\infty} \delta_{n+1}^2 \|\tilde{\eta}_n\|^2 \;,$$

with $\theta^\star \in \arg\min_{\theta \in \Theta} f(\theta)$. $\sum_{n=0}^{+\infty} \delta_{n+1}^2 \|\tilde{\eta}_n\|^2 < +\infty$ almost surely by Lemma 13 since $\sum_{n \in \mathbb{N}} \delta_{n+1}^2 < +\infty$. Since $(\langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)) - \theta^\star, \tilde{\eta}_n^{(a)} \rangle)_{n \in \mathbb{N}}$ is a $(\tilde{\mathcal{F}}_n)_{n \in \mathbb{N}}$-martingale increment, see (21) and by Lemma 14 and $\sum_{n \in \mathbb{N}} \delta_{n+1}^2 / \gamma_n^2 < +\infty$

$$\mathbb{E}\left[\sum_{n=0}^{+\infty} \delta_{n+1}^2 \langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)) - \theta^\star, \tilde{\eta}_n^{(a)} \rangle^2 \right] < +\infty \;,$$

we obtain using [56, Section 12.5] that $\sum_{n=0}^{+\infty} \delta_{n+1} \langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)) - \theta^\star, \tilde{\eta}_n^{(a)} \rangle$ converges almost surely. Using **A**1, (22) and Lemma 16 and Lemma 17 we get that $\sum_{n=0}^{+\infty} \delta_{n+1} \langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)) - \theta^\star, \tilde{\eta}_n^{(i)} \rangle$ is absolutely convergent almost surely for $i \in \{c, d\}$. Finally $\sum_{n=0}^{+\infty} \delta_{n+1} \langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)) - \theta^\star, \tilde{\eta}_n^{(b)} \rangle$ converges almost surely by Lemma 15-(b).

(b) The proof of is identical to the one of Theorem 1-(b).

$\square$

C.4 Proof of Theorem 4

The proof is similar to the one of Theorem 2, using Lemma 13, Lemma 15, Lemma 16, Lemma 17 and the fact that $\tilde{\eta}_n^{(a)}$ is a $(\tilde{\mathcal{F}}_n)_{n\in\mathbb{N}}$-martingale increment, see (21).

C.5 Proof of Theorem 5

In this section, we give the proof of Theorem 5 by showing that **H**1 holds. First of all in Appendix C.5.1, we establish under **L**1 and **L**2 stability results uniform in the parameter $\theta \in \Theta$ for the Langevin diffusion (5) and the associated Euler-Maruyama discretization (6) based on a Foster-Lyapunov drift condition with constants independent of θ. Then, in Appendix C.5.2, we show that the stability conditions that we derive, are sufficient to prove that **H**1 holds. The proof of Theorem 5 then consists in combining all these results and is presented in Appendix C.5.3.

Under **L**1 and **L**2, for any $\theta \in \Theta$, (5) defines a Markov semi-group $(\mathrm{P}_{t,\theta})_{t\geq 0}$ for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $\mathrm{P}_{t,\theta}(x,\mathsf{A}) = \mathbb{P}(Y_t^\theta \in \mathsf{A})$ where $(Y_t^\theta)_{t\geq 0}$ is the solution of (5) with $Y_0^\theta = x$. Consider now the generator of $(\mathrm{P}_{t,\theta})_{t\geq 0}$ for any $\theta \in \Theta$, defined for any $f \in \mathrm{C}^2(\mathbb{R}^d)$ by

$$\mathcal{A}_\theta f = -\langle \nabla_x f, \nabla_x U_\theta(x)\rangle + \Delta_x f \ . \tag{56}$$

We say that a Markov kernel R on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ satisfies a discrete Foster-Lyapunov drift condition $\mathbf{D}_{\mathrm{d}}(V,\lambda,b)$ if there exist $\lambda \in (0,1)$, $b \geq 0$ and a measurable function $V : \mathbb{R}^d \to [1,+\infty)$ such that for all $x \in \mathbb{R}^d$

$$\mathrm{R}V(x) \leq \lambda V(x) + b \ .$$

We say that a Markov semi-group $(\mathrm{P}_t)_{t\geq 0}$ on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ with extended infinitesimal generator $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$ (see e.g. [42] for the definition of $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$) satisfies a continuous drift condition $\mathbf{D}_{\mathrm{c}}(V,\zeta,\beta)$ if there exist $\zeta > 0$, $\beta \geq 0$ and a measurable function $V : \mathbb{R}^d \to [1,+\infty)$ with $V \in \mathrm{D}(\mathcal{A})$ such that for all $x \in \mathbb{R}^d$

$$\mathcal{A}V(x) \leq -\zeta V(x) + \beta \ .$$

C.5.1 Foster-Lyapunov drift conditions uniform on θ

Define $V_{\mathrm{e}} : \mathbb{R}^d \to [1,+\infty)$ for all $x \in \mathbb{R}^d$ by

$$V_{\mathrm{e}}(x) = \exp(\tilde{\mathtt{m}}_1 \phi(x)) \ , \qquad \text{with } \phi(x) = \sqrt{1+\|x\|^2} \text{ and } \tilde{\mathtt{m}}_1 = \mathtt{m}_1/4 \ . \tag{57}$$

Proposition 18. *Assume* **L**1 *and* **L**2*. Let* $\bar{\gamma} < \min(1, 2\mathtt{m}_2)$*. Then there exist* $\lambda_{\mathrm{e}} \in (0,1)$ *and* $b_{\mathrm{e}} \geq 0$ *such that for all* $\gamma \in (0,\bar{\gamma}]$ *and* $\theta \in \Theta$ *the Markov kernel* $\mathrm{R}_{\gamma,\theta}$ *associated with the recursion* (6) *satisfies the discrete drift condition* $\mathbf{D}_{\mathrm{d}}(V,\lambda^\gamma,b\gamma)$*, i.e. for all* $x \in \mathbb{R}^d$

$$\mathrm{R}_{\gamma,\theta} V_{\mathrm{e}}(x) \leq \lambda_{\mathrm{e}}^\gamma V_{\mathrm{e}}(x) + b_{\mathrm{e}}\gamma \mathbb{1}_{\mathrm{B}(0,r_{\mathrm{e}})}(x) \ , \tag{58}$$

with

$$\lambda_{\mathrm{e}} = \mathrm{e}^{-\tilde{\mathtt{m}}_1^2(2^{1/2}-1)} \ , \quad r_{\mathrm{e}} = \max(1, 2(d+\mathtt{c})/\mathtt{m}_1, R_1) \ ,$$
$$b_{\mathrm{e}} = \tilde{\mathtt{m}}_1(d+c+2^{1/2}\tilde{\mathtt{m}}_1)\exp\left[\tilde{\mathtt{m}}_1\left\{(d+c+\tilde{\mathtt{m}}_1)\bar{\gamma} + \sqrt{1+r_{\mathrm{e}}^2}\right\}\right] \ .$$

Proof. Since ϕ is 1-Lipschitz, by the log-Sobolev inequality [4, Proposition 5.4.1], we have for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$\begin{aligned} \mathrm{R}_{\gamma,\theta} V_{\mathrm{e}}(x) &\leq \exp\left[\tilde{\mathtt{m}}_1 \mathrm{R}_{\gamma,\theta}\phi(x) + \tilde{\mathtt{m}}_1^2\gamma\right] \\ &\leq \exp\left[\tilde{\mathtt{m}}_1 \sqrt{\|x - \gamma\nabla_x U_\theta(x)\|^2 + 2\gamma d + 1} + \tilde{\mathtt{m}}_1^2\gamma\right] , \end{aligned} \tag{59}$$

where we have used Jensen's inequality in the last line. Second, using **L**2 and $\gamma < 2\mathtt{m}_2$, we obtain that for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$\begin{aligned} \|x - \gamma\nabla_x U_\theta(x)\|^2 &\leq \|x\|^2 - 2\gamma\langle x, \nabla_x U_\theta(x)\rangle + \gamma^2\|\nabla_x U_\theta(x)\|^2 \\ &\leq \|x\|^2 - 2\mathtt{m}_1\gamma\|x\|\mathbb{1}_{\mathrm{B}(0,R_1)^{\mathrm{c}}}(x) + \gamma(\gamma - 2\mathtt{m}_2)\|\nabla_x U_\theta(x)\|^2 + 2\gamma\mathtt{c} \\ &\leq \|x\|^2 - 2\mathtt{m}_1\gamma\|x\|\mathbb{1}_{\mathrm{B}(0,R_1)^{\mathrm{c}}}(x) + 2\gamma\mathtt{c} . \end{aligned}$$

Therefore, using for any $a > 0$, $\sqrt{1+a} - 1 \leq a/2$, we get for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$\begin{aligned} &\sqrt{\|x - \gamma\nabla_x U_\theta(x)\|^2 + 2\gamma d + 1} - \phi(x) \\ &\leq \phi(x)\left\{\sqrt{1 + 2\gamma\phi^{-2}(x)(d + \mathtt{c} - \mathtt{m}_1\|x\|\mathbb{1}_{\mathrm{B}(0,R_1)^{\mathrm{c}}}(x))} - 1\right\} \\ &\leq \gamma\phi^{-1}(x)(d + \mathtt{c} - \mathtt{m}_1\|x\|\mathbb{1}_{\mathrm{B}(0,R_1)^{\mathrm{c}}}(x)) . \end{aligned} \tag{60}$$

Therefore, combining this result with (59) and using that for any $\tilde{x} \in \overline{\mathrm{B}}(0, r_{\mathrm{e}})^{\mathrm{c}}$, $\phi(\tilde{x})^2/\|\tilde{x}\|^2 \leq 2$ and $d + c \leq \mathtt{m}_1\|x\|/2$, we obtain for any $x \in \overline{\mathrm{B}}(0, r_{\mathrm{e}})^{\mathrm{c}}$ and $\theta \in \Theta$,

$$\begin{aligned} \mathrm{R}_{\gamma,\theta} V_{\mathrm{e}}(x) &\leq \exp\left[\tilde{\mathtt{m}}_1\phi^{-1}(x)(d + \mathtt{c} - \mathtt{m}_1\|x\|) + \tilde{\mathtt{m}}_1^2\gamma\right] V_{\mathrm{e}}(x) \\ &\leq \exp\left[-2\tilde{\mathtt{m}}_1^2\gamma\phi^{-1}(x)\|x\| + \tilde{\mathtt{m}}_1^2\gamma\right] V_{\mathrm{e}}(x) \leq \lambda_{\mathrm{e}}^{\gamma} V_{\mathrm{e}}(x) . \end{aligned}$$

Using (59), (60), and the fact that $\phi(\tilde{x}) \geq 1$ for any $\tilde{x} \in \mathbb{R}^d$, we have for any $x \in \mathrm{B}(0, r_{\mathrm{e}})$ and $\theta \in \Theta$,

$$\mathrm{R}_{\gamma,\theta} V_{\mathrm{e}}(x) \leq \lambda_{\mathrm{e}}^{\gamma} V_{\mathrm{e}}(x) + \left(\mathrm{e}^{\tilde{\mathtt{m}}_1(d+c+\tilde{\mathtt{m}}_1)\gamma} - \lambda_{\mathrm{e}}^{\gamma}\right)\exp\left[\tilde{\mathtt{m}}_1\sqrt{1 + r_{\mathrm{e}}^2}\right] .$$

The proof of (58) for $x \in \mathrm{B}(0, r_{\mathrm{e}})$ and $\theta \in \Theta$ is then completed upon using that $\mathrm{e}^a - \mathrm{e}^b \leq (a - b)\mathrm{e}^a$ for all $a, b \in \mathbb{R}$ with $a \geq b$. □

Proposition 19. *Assume* **L**1 *and* **L**2*. Then for any $\theta \in \Theta$, $(\mathrm{P}_{t,\theta})_{t\geq 0}$ associated with (5) satisfies the continuous drift condition $\mathbf{D}_{\mathrm{c}}(V_{\mathrm{e}}, \zeta_{\mathrm{e}}, \beta_{\mathrm{e}})$ for V_{e} defined in (57) and*

$$\zeta_{\mathrm{e}} = 3\tilde{\mathtt{m}}_1^2/2^{1/2} , \quad \beta_{\mathrm{e}} = \tilde{\mathtt{m}}_1\exp\left[\tilde{\mathtt{m}}_1\sqrt{1 + \tilde{r}_{\mathrm{e}}^2}\right](1 + \tilde{\mathtt{m}}_1 + \mathtt{c} + d) , \quad \tilde{r}_{\mathrm{e}} = \max(1, R_1) .$$

Proof. First, by definition, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned} \nabla_x V(x) &= \tilde{\mathtt{m}}_1 x V(x)/\phi(x) \\ \Delta_x V(x) &= \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\}\{\tilde{\mathtt{m}}_1\|x\|^2/\phi(x) + d - \|x\|^2/\phi^2(x)\} . \end{aligned}$$

Therefore, by (56) and **L**2, we get for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$,

$$\begin{aligned}\mathcal{A}_\theta V(x) &= \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\} \left[-\langle \nabla_x U_\theta(x), x\rangle + \tilde{\mathtt{m}}_1 \|x\|^2/\phi(x) + d - \|x\|^2/\phi^2(x) \right] \\ &\le \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\} \left[-\mathtt{m}_1 \|x\| \mathbb{1}_{\mathrm{B}(0,R_1)^{\mathrm{c}}}(x) + \mathtt{c} + \tilde{\mathtt{m}}_1 \|x\|^2/\phi(x) + d - \|x\|^2/\phi^2(x) \right] \\ &\le \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\} \left[-(3\mathtt{m}_1/4) \|x\| \mathbb{1}_{\mathrm{B}(0,R_1)^{\mathrm{c}}}(x) + \mathtt{c} + \tilde{\mathtt{m}}_1 \|x\| \mathbb{1}_{\mathrm{B}(0,R_1)}(x) + d \right] .\end{aligned}$$

The proof is then complete upon using that for any $x \in \mathrm{B}(0, \tilde{r}_{\mathrm{e}})^{\mathrm{c}}$, $\|x\|/\phi(x) \ge 2^{-1/2}$, for any $y \in \mathbb{R}^d$, $\|y\|/\phi(y) \le 1$.

□

C.5.2 Checking H1

Lemma 20. *Assume* **L**1 *and let* $V : \mathbb{R}^d \to [1, +\infty)$ *satisfying* $\lim_{\|x\|\to+\infty} V(x) = +\infty$ *and* $V \in \mathrm{D}(\mathcal{A}_\theta)$*, for any* $\theta \in \Theta$*, where* $\mathcal{A}_\theta$ *is defined by* (56)*. .*

(a) Assume that there exist $\lambda \in (0,1)$*,* $b \ge 0$ *and* $\bar{\gamma} > 0$ *such that for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$*,* $\mathrm{R}_{\gamma,\theta}$ *associated with the recursion* (24)*, satisfies* $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$*. Then for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$*,* $\mathrm{R}_{\gamma,\theta}$ *admits an invariant probability measure* $\pi_{\gamma,\theta}$ *on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *and there exists* $D_3 \ge 0$ *such that for any* $x \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$

$$\delta_x \mathrm{R}^k_{\gamma,\theta} V \le D_3 + V(x) , \qquad \pi_{\gamma,\theta}(V) \le D_3 , \qquad D_3 = b\lambda^{-\bar{\gamma}}/\log(1/\lambda) .$$

In addition, for all $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\lim_{k\to+\infty} \|\delta_x \mathrm{R}^k_{\gamma,\theta} - \pi_{\gamma,\theta}\|_V = 0$*.*

(b) Assume that there exist $\zeta > 0$ *and* $\beta \ge 0$ *such that for any* $\theta \in \Theta$*,* $(\mathrm{P}_{t,\theta})_{t\ge0}$ *associated with* (5) *satisfies* $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$*. Then for any* $\theta \in \Theta$*, the diffusion is non-explosive,* $\mathcal{A}_\theta$ *admits* π_θ *as an invariant probability measure and*

$$\pi_\theta(V) \le D_0 , \qquad D_0 = \beta/\zeta .$$

In addition, for all $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\lim_{t\to+\infty} \|\delta_x \mathrm{P}_{t,\theta} - \pi_\theta\|_V = 0$*.*

Proof. (a) for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, $\mathrm{R}_{\gamma,\theta}$ is irreducible with respect to the Lebesgue measure on $\mathbb{R}^d$, has the Feller property and satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$ then [41, Section 4.4] applies and $\mathrm{R}_{\gamma,\theta}$ admits an invariant probability measure $\pi_{\gamma,\theta}$. The discrete drift condition and [21, Lemma 1] give that for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$

$$\mathrm{R}^k_{\gamma,\theta} V(x) \le V(x) + b\lambda^{-\bar{\gamma}}/\log(1/\lambda) , \qquad \pi_{\gamma,\theta}(V) \le b\lambda^{-\bar{\gamma}}/\log(1/\lambda) .$$

We obtain that for all $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\lim_{k\to+\infty} \|\delta_x \mathrm{R}^k_{\gamma,\theta} - \pi_{\gamma,\theta}\|_V = 0$ using [39, Theorem 16.0.1].

(b) Using $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$ and [42, Theorem 2.1] we get that the diffusion process is non-explosive and thus $(\mathrm{P}_{t,\theta})_{t\ge0}$ is defined for any $\theta \in \Theta$ and $t \ge 0$. Using [52, Corollary 10.1.4] for any $\theta \in \Theta$, $(\mathrm{P}_{t,\theta})_{t\ge0}$ is strongly Feller continuous, therefore any compact sets is petite for the Markov kernel $\mathrm{P}_{h,\theta}$, for any $h > 0$ and $\theta \in \Theta$, by [39, Theorem 6.0.1]. Using [47, Chapter 7, Proposition 1.5], [27, Chapter 4, Theorem 9.17], and the fact that $\pi_\theta(\mathcal{A}_\theta f) = 0$ for any $\theta \in \Theta$ and $f \in \mathrm{C}^2_c(\mathbb{R}^d)$, we obtain that for any $\theta \in \Theta$, π_θ is an invariant measure for $(\mathrm{P}_{t,\theta})_{t\ge0}$. Using $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$ and [42, Theorem 4.5] we get that for all $\theta \in \Theta$, $\pi_\theta(V) \le \beta/\zeta$. Finally, the convergence is ensured using [40, Theorem 5.1].

□

As an immediate corollary we obtain that under the conditions of Lemma 20 for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $k \in \mathbb{N}$,

$$\pi_\theta \mathrm{R}^k_{\gamma,\theta} V \leq \beta/\zeta + b\lambda^{-\bar{\gamma}}/\log(1/\lambda) \; . \tag{61}$$

Lemma 21. *Let $V : \mathbb{R}^d \to [1, +\infty)$. Assume there exist $\lambda \in (0,1)$, $b \geq 0$ and $\bar{\gamma} > 0$ such that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, $\mathrm{R}_{\gamma,\theta}$ associated with the recursion (6) satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of positive integers satisfying $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$. Then, $(X^n_k)_{n\in\mathbb{N}, k\in\{0,\ldots,m_n\}}$ given by (14) with $\{\mathrm{K}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{\mathrm{R}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ satisfies for all $p, n \in \mathbb{N}$ and $k \in \{0, \ldots, m_n\}$*

$$\mathbb{E}\left[\mathrm{R}^p_{\gamma_n,\theta_n} V(X^n_k) \middle| X^0_0 \right] \leq D_1 V(X^0_0) \; , \qquad D_1 = 1 + 2b\lambda^{-\bar{\gamma}}/\log(1/\lambda) \; .$$

Proof. By induction we obtain that

$$\mathbb{E}\left[V(X^{n+1}_k) \middle| \mathcal{F}_n \right] = \mathrm{R}^k_{\gamma_{n+1},\theta_{n+1}} V(X^{n+1}_0) \leq \lambda^{k\gamma_{n+1}} V(X^{n+1}_0) + b\gamma_{n+1} \sum_{i=1}^{k} \lambda^{\gamma_{n+1}(k-i)} \; , \tag{62}$$

where $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is defined by (15). Similarly, we obtain for any $k \in \{0, \ldots, m_0\}$,

$$\mathbb{E}\left[V(X^0_k) \middle| X^0_0 \right] = \mathrm{R}^k_{\gamma_0,\theta_0} V(X^0_0) \leq \lambda^{k\gamma_0} V(X^0_0) + b\gamma_0 \sum_{i=1}^{k} \lambda^{\gamma_0(k-i)} \; . \tag{63}$$

Define for $\ell \in \mathbb{N}, k \in \mathbb{N}$ and $i \in \mathbb{N}^*$, $\mathrm{q}_{\ell,k} = \sum_{j=0}^{\ell-1} m_j + k$, $\mathrm{q}_n = \mathrm{q}_{\ell,0}$ and $\tilde{\gamma}_i = \sum_{j=0}^{+\infty} \gamma_j \mathbb{1}_{(\mathrm{q}_j, \mathrm{q}_{j+1}]}(i)$. In addition, consider for any $\mathrm{p}, \mathrm{q} \in \mathbb{N}^*$, $\Gamma_{\mathrm{p},\mathrm{q}} = \sum_{i=\mathrm{p}}^{\mathrm{q}} \tilde{\gamma}_i$ and $\Gamma_{\mathrm{p}} = \Gamma_{1,\mathrm{p}}$. Combining (62), (63) and Lemma 9 we get for any $n \in \mathbb{N}$ and $k \in \{0, \ldots, m_n\}$

$$\begin{aligned} \mathbb{E}\left[\mathrm{R}^p_{\gamma_n,\theta_n} V(X^n_k) \middle| X^0_0 \right] &\leq \lambda^{\gamma_n p} \mathbb{E}\left[V(X^n_k) \middle| X^0_0 \right] + b\log(1/\lambda)\lambda^{-\bar{\gamma}} \\ &\leq \lambda^{\Gamma_{\mathrm{q}_{n,k}}} V(X^0_0) + b \sum_{i=1}^{\mathrm{q}_{n,k}} \tilde{\gamma}_i \lambda^{\Gamma_{i+1,\mathrm{q}_{n,k}}} + b\log(1/\lambda)\lambda^{-\bar{\gamma}} \; . \end{aligned} \tag{64}$$

Since $(\tilde{\gamma}_i)_{i\in\mathbb{N}}$ is nonincreasing and for all $t \geq 0$, $1 - \lambda^t \geq -t\lambda^t \log(\lambda)$, we have for all $\mathrm{q} \in \mathbb{N}^*$,

$$\begin{aligned} &\sum_{i=1}^{\mathrm{q}} \tilde{\gamma}_i \lambda^{\Gamma_{i+1,\mathrm{q}}} \leq \sum_{i=1}^{\mathrm{q}} \tilde{\gamma}_i \prod_{j=i+1}^{\mathrm{q}} (1 + \lambda^{\tilde{\gamma}_1} \log(\lambda)\tilde{\gamma}_j) \\ &\leq (-\lambda^{\tilde{\gamma}_1} \log(\lambda))^{-1} \sum_{i=1}^{\mathrm{q}} \left\{ \prod_{j=i+1}^{\mathrm{q}} (1 + \lambda^{\tilde{\gamma}_1} \log(\lambda)\tilde{\gamma}_j) - \prod_{j=i}^{\mathrm{q}} (1 + \lambda^{\tilde{\gamma}_1} \log(\lambda)\tilde{\gamma}_j) \right\} \\ &\leq (-\lambda^{\tilde{\gamma}_1} \log(\lambda))^{-1} \; . \end{aligned}$$

Combining this result and (64) completes the proof. □

Lemma 22. *Let $V : \mathbb{R}^d \to [1, +\infty)$ measurable and $M_V \geq 0$ such that $\sup_{x\in\mathbb{R}^d} \left\{ (1 + \|x\|)^2 / V(x) \right\} \leq M_V$. Assume* **L1** *and that for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $k \in \mathbb{N}$,*

$$\pi_\theta \mathrm{R}^k_{\gamma,\theta}(V) \leq \tilde{D}_1 \; , \qquad \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} V \leq \tilde{D}_1 \; , \tag{65}$$

with $m_\gamma = \lceil 1/\gamma \rceil$. *Then for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$

$$\|\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} - \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\|_{V^{1/2}}^2 \\ \leq 2\tilde{D}_1 \mathtt{L}^2 \gamma(1+\bar{\gamma}) \left\{ d + 2\bar{\gamma}(\sup_{\theta\in\Theta} \|\nabla_x U_\theta(0)\|^2 + \mathtt{L}^2 M_V \tilde{D}_1) \right\} ,$$

Proof. The proof follows the lines of [21, Theorem 10]. Let $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$. We have, using a generalized Pinsker inequality [21, Lemma 24], that

$$\begin{aligned} \|\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} - \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\|_{V^{1/2}}^2 &\leq 2(\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} V + \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} V) \mathrm{KL}\left(\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} | \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\right) . \\ &\leq 4\tilde{D}_1 \mathrm{KL}\left(\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} | \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\right) . \end{aligned}$$

Using **L** 1, [21, Equation (15)], [36, Theorem 4.1, Chapter 2], (65) and that for any $a, b \in \mathbb{R}$, $(a+b)^2 \leq 2(a^2+b^2)$ we obtain that

$$\begin{aligned} \mathrm{KL}\left(\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} | \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\right) &\leq \mathtt{L}^2 m_\gamma \gamma^2 (d + \bar{\gamma} \sup_{k\in\mathbb{N}} \pi_\theta \mathrm{R}_{\gamma,\theta}^k \|\nabla_x U_\theta(x)\|^2) \\ &\leq \mathtt{L}^2 (1+\bar{\gamma})\gamma(d + 2\bar{\gamma}(\sup_{\theta\in\Theta} \|\nabla_x U_\theta(0)\|^2 + \mathtt{L}^2 M_V \tilde{D}_1)) , \end{aligned}$$

which concludes the proof. □

Proposition 23. *Let* $V : \mathbb{R}^d \to [1, +\infty)$ *measurable and* $M_V \geq 0$ *such that* $\sup_{x\in\mathbb{R}^d} \left\{(1+\|x\|)^2/V(x)\right\} \leq M_V$. *Assume* **L** 1 *and that there exist* $\lambda \in (0,1)$, $b \geq 0$ *and* $\bar{\gamma} > 0$ *such that for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$ $\mathrm{R}_{\gamma,\theta}$ *satisifies* $\mathbf{D}_\mathrm{d}(V, \lambda^\gamma, b\gamma)$. *Assume that there exists* $D_0 \geq 0$ *such that for any* $\theta \in \Theta$, $\pi_\theta(V) \leq D_0$. *Then there exists* $D_4 \geq 0$ *such that for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$

$$\|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \leq D_4 \gamma^{1/2} .$$

Proof. Using Lemma 20 we obtain that for any $\theta \in \Theta$

$$\lim_{k\to+\infty} \|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} = \|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} . \tag{66}$$

We now give an upper bound on $\|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}}$ for $k = q_\gamma m_\gamma$ with $m_\gamma = \lceil 1/\gamma \rceil$ and $q_\gamma \in \mathbb{N}$. Using [16, Theorem 6] and that π_θ is invariant for $\mathrm{P}_{t,\theta}$ with $t \geq 0$, see Lemma 20, we obtain for all $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $k \in \mathbb{N}$

$$\begin{aligned} &\|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} \\ &\leq \sum_{\ell=0}^{q_\gamma-1} \|\pi_\theta \mathrm{P}_{\gamma(\ell+1)m_\gamma,\theta} \mathrm{R}_{\gamma,\theta}^{(q_\gamma-(\ell+1))m_\gamma} - \pi_\theta \mathrm{P}_{\gamma \ell m_\gamma,\theta} \mathrm{R}_{\gamma,\theta}^{(q_\gamma-\ell)m_\gamma}\|_{V^{1/2}} \\ &\leq \sum_{\ell=0}^{q_\gamma-1} C\xi^{\gamma m_\gamma(q_\gamma-(\ell+1))} \|\pi_\theta \mathrm{P}_{\gamma \ell m_\gamma,\theta} \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{P}_{\gamma \ell m_\gamma,\theta} \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \\ &\leq \|\pi_\theta \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \sum_{\ell=1}^{q_\gamma} C\xi^{\ell\gamma m_\gamma} , \end{aligned} \tag{67}$$

where $C \geq 0, \xi \in (0,1)$ are the constants given by [16, Theorem 6] with minorization condition given by [16, Proposition 8a] with $\mathtt{m} = -\mathtt{L}$ since $\mathbf{L}1$ holds and drift condition $\mathbf{D}_{\mathrm{d}}(V^{1/2}, \lambda^{\gamma}, b\lambda^{-\bar{\gamma}/2}\gamma/2)$, since for all $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ we have that $\mathrm{R}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^{\gamma}, b\gamma)$ and therefore using Jensen's inequality that $\mathrm{R}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V^{1/2}, \lambda^{\gamma/2}, b\lambda^{-\bar{\gamma}/2}\gamma/2)$.

We now give an upper bound on error $\|\pi_\theta \mathrm{P}_{m_\gamma \gamma,\theta} - \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}}$. Indeed, since $\mathcal{A}_\theta$ satisfies a $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$ and $\mathrm{R}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^{\gamma}, b\gamma)$ for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, we obtain using (61) that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$

$$\pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}(V) \leq D_0 , \qquad \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}(V) \leq \tilde{D}_1 , \qquad \tilde{D}_1 = D_0 + b\lambda^{-\bar{\gamma}} \log(1/\lambda)^{-1} ,$$

Combining this result and Lemma 22 we have for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$

$$\|\pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} - \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \leq \tilde{D}_2 \gamma^{1/2} , \tag{68}$$

with

$$\tilde{D}_2 = 2\tilde{D}_1^{1/2}(1+\bar{\gamma})^{1/2} \left\{ d + 2\bar{\gamma}(\mathtt{L}^2 M_V + \sup_{\theta\in\Theta} \|\nabla_x U_\theta(0)\|^2)\tilde{D}_1 \right\}^{1/2} \mathtt{L} .$$

Combining (67) and (68) we get for any $k \in \mathbb{N}$, $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$

$$\|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} \leq C\tilde{D}_2 \sum_{\ell=1}^{q_\gamma} \xi^{\gamma m_\gamma \ell} \gamma^{1/2} \leq C\tilde{D}_2 (1-\xi)^{-1} \gamma^{1/2} ,$$

where we used that $\xi^{\gamma m_\gamma} \leq \xi$. The conclusion follows from this result and (66). $\square$

C.5.3 Proof of Theorem 5

Combining Proposition 18 and Lemma 21 we get that $\mathbf{H}1$-(i) is satisfied with constant $A_1 \leftarrow D_1$. $\mathbf{L}1$, $\mathbf{L}2$, Proposition 18 and Lemma 20-(a) ensure that $\mathbf{H}1$-(ii) is satisfied by [16, Theorem 14] with $A_3 \leftarrow D_3$. $\mathbf{H}1$-(iii) is satisfied combining Proposition 18, Proposition 19 and Proposition 23 with $\mathbf{\Psi}(\gamma) \leftarrow D_4 \gamma^{1/2}$.

C.6 Proof of Theorem 6

We preface the proof by a technical lemma.

Proposition 24. *Let $V : \mathbb{R}^d \to [1, +\infty)$ and $M_{V,4} \geq 0$ such that $\sup_{x\in\mathbb{R}^d} \{(1 + \|x\|^4)/V(x)\} \leq M_{V,4}$. Assume that there exists $M \geq 1$ such that for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$, with $\bar{\gamma} > 0$ and $x \in \mathbb{R}^d$, $\mathrm{R}_{\gamma,\theta} V(x) \leq MV(x)$. Assume $\mathbf{L}1$ and $\mathbf{L}3$, then we have for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_2 < \gamma_1$, $a \in [1/4, 1/2]$ and $x \in \mathbb{R}^d$*

$$\|\delta_x \mathrm{R}_{\gamma_1,\theta_1} - \delta_x \mathrm{R}_{\gamma_2,\theta_2}\|_{V^a} \leq D_5 \left[\gamma_1/\gamma_2 - 1 + \gamma_2^{1/2} \|\theta_1 - \theta_2\| \right] V(x)^{2a} ,$$

where $\{\mathrm{R}_{\gamma,\theta}, \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ is the sequence of Markov kernels associated with the recursion (6) and

$$D_5 = \max \left(2M^{1/2} \left[d/4 + \sup_{\theta\in\Theta} \|\nabla_x U_\theta(0)\|^2 + \mathtt{L}^2 M_{4,V}^{1/2} \right]^{1/2} , (2M)^{1/2} L_U \right) .$$

Proof. Let $x \in \mathbb{R}^d$, $\theta_1, \theta_2 \in \Theta$ and $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$, $\gamma_2 < \gamma_1$. Using [21, Lemma 24] we have that

$$
\begin{aligned}
&\|\delta_x \mathrm{R}_{\gamma_1,\theta_1} - \delta_x \mathrm{R}_{\gamma_2,\theta_2}\|_{V^a} \\
&\quad \leq \sqrt{2}\left(\mathrm{R}_{\gamma_1,\theta_1} V^{2a}(x) + \mathrm{R}_{\gamma_2,\theta_2} V^{2a}(x)\right)^{1/2} \mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right)^{1/2} \\
&\quad \leq 2M^a V^a(x) \mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right)^{1/2}
\end{aligned} \tag{69}
$$

Denote for any $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$, γ_{μ,σ^2} the d-dimensional Gaussian distribution with mean μ and covariance matrix $\sigma^2 \,\mathrm{Id}$. Using that for any $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma_1, \sigma_2 > 0$,

$$
\mathrm{KL}\left(\Upsilon_{\mu_1,\sigma_1 \mathrm{Id}} | \Upsilon_{\mu_2,\sigma_2 \mathrm{Id}}\right) = \|\mu_1 - \mu_2\|^2 / (2\sigma_2^2) + (d/2)\left\{-\log(\sigma_1^2/\sigma_2^2) - 1 + \sigma_1^2/\sigma_2^2\right\} .
$$

In addition, if $\sigma_1 \geq \sigma_2$

$$
\mathrm{KL}\left(\Upsilon_{\mu_1,\sigma_1 \mathrm{Id}} | \Upsilon_{\mu_2,\sigma_2 \mathrm{Id}}\right) \leq \|\mu_1 - \mu_2\|^2 / (2\sigma_2^2) + (d/2)(1 - \sigma_1^2/\sigma_2^2)^2 .
$$

Therefore, we obtain that

$$
\mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right) \leq \Xi/(4\gamma_2) + (d/2)(1 - \gamma_1/\gamma_2)^2 , \tag{70}
$$

where Ξ satisfies

$$
\begin{aligned}
\Xi &= \|\gamma_1 \nabla_x U_{\theta_1}(x) - \gamma_2 \nabla_x U_{\theta_2}(x)\|^2 \\
&= \|\gamma_1 \nabla_x U_{\theta_1}(x) - \gamma_2 \nabla_x U_{\theta_1}(x) + \gamma_2 \nabla_x U_{\theta_1}(x) - \gamma_2 \nabla_x U_{\theta_2}(x)\|^2 \\
&\leq 2\|\gamma_1 \nabla_x U_{\theta_1}(x) - \gamma_2 \nabla_x U_{\theta_1}(x)\|^2 + 2\|\gamma_2 \nabla_x U_{\theta_1}(x) - \gamma_2 \nabla_x U_{\theta_2}(x)\|^2 \\
&\leq 2(\gamma_1 - \gamma_2)^2 \|\nabla_x U_{\theta_1}(x)\|^2 + 2\gamma_2^2 \|\nabla_x U_{\theta_1}(x) - \nabla_x U_{\theta_2}(x)\|^2 \\
&\leq 2(\gamma_1 - \gamma_2)^2 \|\nabla_x U_{\theta_1}(x)\|^2 + 2\gamma_2^2 L_U^2 \|\theta_1 - \theta_2\|^2 V^{2a}(x) ,
\end{aligned} \tag{71}
$$

where we have used **L**3 in the last line. Using **L**3 again and that $\sup_{\theta \in \Theta} \|\nabla_x U_\theta(0)\| < +\infty$ by **L**1, we get for any $a \in [1/4, 1/2]$

$$
\|\nabla_x U_\theta(x)\|^2 \leq 2(\|\nabla_x U_\theta(x) - \nabla_x U_\theta(0)\|^2 + \sup_{\theta \in \Theta} \|\nabla_x U_\theta(0)\|^2) \leq C_\Theta V^{2a}(x) ,
$$

with $C_\Theta = 2\sup_{\theta \in \Theta} \|\nabla_x U_\theta(0)\|^2 + 2\mathrm{L}^2 M_{4,V}^{1/2}$. Combining this result, $\log(\gamma_2/\gamma_1) \leq 0$ and and (71) in (70), it follows that

$$
\begin{aligned}
&\mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right) \leq d(1 - \gamma_1/\gamma_2)^2/2 \\
&+ \gamma_2^{-1}(\gamma_1 - \gamma_2)^2 \|\nabla_x U(\theta_1, x)\|^2/2 + \gamma_2 L_U^2 \|\theta_1 - \theta_2\|^2 V^{2a}(x)/2 \\
&\leq \left[d\gamma_2^{-1}(1 - \gamma_2/\gamma_1)/4 + \gamma_2^{-1}(\gamma_1 - \gamma_2)^2 C_\Theta/2 + \gamma_2 L_U^2 \|\theta_1 - \theta_2\|^2/2\right] V^{2a}(x) .
\end{aligned}
$$

This result substituted in (69) completes the proof with the fact that for any $a, b \in \mathbb{R}_+$, $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$. □

Proof of Theorem 6. **L**1 and **L**2 ensure a uniform drift condition on $\mathrm{R}_{\gamma,\theta}$, see Proposition 18 . Note that the Lyapunov function V defined by Proposition 18 satisfies $\sup_{x \in \mathbb{R}^d}(1 + \|x\|^4)/V(x) < +\infty$. **H**2 is then a direct consequence of Proposition 24 □