Statistics and Computing (2021) 31:29
https://doi.org/10.1007/s11222-020-09986-y

Efficient stochastic optimisation by unadjusted Langevin Monte Carlo

Application to maximum marginal likelihood and empirical Bayesian estimation

Valentin De Bortoli[1,3] · Alain Durmus[1] · Marcelo Pereyra[2] · Ana F. Vidal[2]

Received: 26 December 2019 / Accepted: 17 September 2020 / Published online: 19 March 2021

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 statistical audio analysis and a sparse Bayesian logistic regression with random effects problem.

Keywords Maximum marginal likelihood estimation · Empirical Bayesian inference · Stochastic approximation · Markov chain Monte Carlo methods · Unadjusted Langevin Algorithm · Recursive estimation

1 Introduction

Maximum likelihood estimation (MLE) is central to modern statistical science. It is a cornerstone of frequentist inference (Berger and Casella 2002), and also plays a fundamental role in parametric empirical Bayesian inference (Carlin and Louis 2000; Casella 1985). For simple statistical models, MLE can be performed analytically and exactly. 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 (Gentle et al. 2012; Boyd and Vandenberghe 2004).

In this paper, we consider MLE in models involving incomplete or "missing" data, such as hidden, latent or unobserved variables. Expectation maximisation (EM) optimisation methods (Dempster et al. 1977) are common strategies to obtain approximate solutions in this setting. However, they rely on a maximization step of a surrogate which is not possible in most models (Robert and Casella 2004). Several strategies can be considered to overcome this issue. In particular, we consider Robbins–Monro stochastic approximation (SA) schemes that use a Monte Carlo stochastic simulation algorithm to approximate the gradients that drive the optimisation procedure (Robbins and Monro 1951; Delyon et al. 1999; Kushner and Yin 2003; Fort et al. 2011). When combined with Markov chain Monte Carlo (MCMC) algorithm, SA schemes provide a powerful general methodology which

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-020-09986-y.

✉ Valentin De Bortoli
valentin.debortoli@gmail.com

[1] CMLA, ENS Paris-Saclay & CNRS, 61 Avenue du Président Wilson, 94230 Cachan, France

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

[3] Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK

is simple to implement, has a detailed convergence theory (Atchadé et al. 2017), and can be easily applied to a wide range of moderately low-dimensional models.

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. For large models, these are usually high-dimensional Metropolis–Hastings methods such as the Metropolis-adjusted Langevin algorithm (Roberts and Tweedie 1996) or Hamiltonian Monte Carlo (Girolami and Calderhead 2011; Durmus et al. 2017), which are sometimes difficult to calibrate within the SA scheme to achieve a prescribed acceptance rate (both automatic and manual calibration can be difficult as the target density changes with each iteration of the SA scheme). For these reasons, practitioners rarely use SA schemes with Markovian disturbances in high-dimensional settings.

In this paper, we propose to address these limitations by exploiting recent developments in inexact MCMC methodology to drive the SA scheme, particularly unadjusted Langevin algorithms, which have easily verifiable geometric ergodicity conditions and are easy to calibrate (Durmus and Moulines 2017; Dalalyan 2017). 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: Sect. 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. Detailed theoretical convergence results for the method are reported in Sect. 3, which also describes a generalisation of the proposed methodology and theory to other inexact Markov kernels. Section 4 presents three numerical experiments that demonstrate the proposed methodology in a variety of scenarios. The online supplementary material includes additional theoretical results, postponed proofs 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

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

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 (Dempster et al. 1977). It is central to parametric empirical Bayes approaches that base their inferences on the pseudo-posterior distribution given by $p(x|y, \theta^\star) = p(y, x|\theta^\star)/p(y|\theta^\star)$ (Carlin and Louis 2000; Vidal et al. 2019). The same 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 $g(\theta) = -\log p(\theta)$ (Berger and Casella 2002).

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 Vidal et al. (2019), De Bortoli et al. (2020b) which focus 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 (Delyon et al. 1999), 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

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

where $\bar{H}_{\theta_n}$ 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 (Boyd and Vandenberghe 2004) for models with intractable gradients. For $n \in \mathbb{N}$, Monte Carlo estimators $\bar{H}_{\theta_n}$ for $\nabla_\theta \log p(y|\theta)$ at θ_n are derived from Fisher's 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

$$\bar{H}_{\theta_n}(X^n_{1:m_n}) = m_n^{-1} \sum_{k=1}^{m_n} \nabla_\theta \log p(X^n_k, y|\theta_n) \ , \tag{3}$$

where $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$ is a sequence of batch sizes and $X^n_{1:m_n} = (X^n_k)_{k\in\{1,\ldots,m_n\}}$ is either a sequence of exact Monte Carlo samples from $p(x|y, \theta_n) = p(x, y|\theta_n)/p(y|\theta_n)$, or a Markov chain targeting this distribution.

Given a sequence $(\theta_k)_{k\in\{1,\ldots,N\}}$ 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 a.s to a solution of (1) as $N \to \infty$ provided that some conditions on $p(y|\theta)$, $p(x|y, \theta)$, g, $(\delta_n)_{n\in\mathbb{N}}$, and $\bar{H}_{\theta_n}$ 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 Benveniste et al. 1990; Fort and Moulines 2003; Duchi et al. 2011; Andrieu and Moulines 2006; Nemirovski et al. 2008; Atchadé et al. 2017).

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 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.

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 lately because they can exhibit excellent large-scale convergence properties and empirically outperform their Metropolised counterparts in many situations (see Durmus et al. 2018 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 (Durmus and Moulines 2017; Dalalyan 2017; Eberle and Majka 2018; De Bortoli and Durmus 2019). 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 $(\boldsymbol{X}^\theta_t)_{t\geq 0}$ for which the target distribution $p(x|y, \theta)$ is invariant. One fundamental example is the Langevin dynamics solution of the following stochastic differential equation (SDE)

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

where $(\boldsymbol{B}_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. Under mild assumptions on $p(x|y, \theta)$, this SDE admits a strong solution for which $p(x|y, \theta)$ is the invariant probability measure. In addition, there are detailed explicit convergence results for $(\boldsymbol{X}^\theta_t)_{t\geq 0}$ to $p(x|y, \theta)$ under different metrics (Eberle 2016; Eberle et al. 2017). Note that other SDEs satisfy these favorable properties such as the kinetic Langevin dynamics (Dalalyan and Riou-Durand 2018).

However, sampling solutions for these continuous-time dynamics is not feasible in general. Therefore, discretisations 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) (Roberts and Tweedie 1996), 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 discretisation time step and $(Z_k)_{k\in\mathbb{N}^*}$ is a i.i.d. sequence of d-dimensional zero-mean Gaussian random variables with identity covariance matrix. 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 (Roberts and Tweedie 1996). This correction empirically 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 Durmus et al. 2018). 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 (Dalalyan 2017; Durmus and Moulines 2017). In addition, new explicit convergence bounds have been derived under various assumptions on the target probability distribution (Dalalyan 2017; Cheng et al. 2018; Cheng and Bartlett 2017; Lee et al. 2018). In addition, accelerations and variations of ULA have been studied, both theoretically and experimentally, yielding better ergodic convergence rates (Maddison et al. 2018; Ma et al. 2019; Muehlebach and Jordan 2019; Dalalyan and Riou-Durand 2018). However, such extensions are out of the scope of the present work whose main contribution is not to provide new results to the existing Markov chain theory but to use the theoretical guarantees of ULA in order to study SA schemes driven by this efficient but inexact sampler.

Note also that the methodology we propose and analyse in this paper is fundamentally different from the Stochastic Gradient Langevin dynamics (Vollmer et al. 2016; Teh et al. 2016; Welling and Teh 2011a; Patterson and Teh 2013; Ahn et al. 2014, 2012) which is an MCMC algorithm to sample from $p(x|y,\theta)$ using estimators of $\nabla_x \log p(x|y,\theta)$. Finally, it should be highlighted that, in an independent line of work, a similar methodology is studied under a different set of assumptions in (Karimi et al. 2019).

2.4 The SOUL algorithm and main results

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 stepsizes 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 associated with (6) to approximately sample from $p(x|y,\theta)$, and by $(\gamma_n)_{n\in\mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ the sequence of discrete time steps used.

Formally, starting from some $X_0^0 \in \mathbb{R}^d$ and $\theta_0 \in \Theta$ we define recursively $(\{X_k^n : k \in \{0,\dots,m_n\}\}, \theta_n)_{n\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$: for any $n \in \mathbb{N}$ $X_{0:m_n}^n$ 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

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

where for any $n \in \mathbb{N}$ and $k \in \{1,\dots,m_n\}$, $\bar{H}_{\theta_n}(X_k^n) = \nabla_\theta \log p(X_k^n, y|\theta_n)$ and we recall that Π_Θ is the projection onto Θ, and for all $n \in \mathbb{N}$

$$\mathcal{F}_n = \sigma\left(\theta_0, \{X_{1:m_\ell}^\ell : \ell \in \{0,\dots,n\}\}\right),$$
$$\mathcal{F}_{-1} = \sigma(\theta_0). \tag{7}$$

Note that such a construction is always possible by Kolmogorov extension theorem (Kallenberg 2006, 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 the average of the iterates (4). 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 $n \in \{1,\dots,N\}$.

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

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

In Sect. 3 we prove the following results for SOUL, which we derive in more generality by analysing a broader class of methods where the Markov kernel associated with ULA is replaced by any regular and geometrically ergodic Markov kernel, see H1 and H2. For any $a \in \mathbb{R}^d$ and $R \geq 0$ denote by $\mathrm{B}(a,R)$ the open ball centered at a and radius $R \geq 0$ and $\overline{\mathrm{B}}(a,R)$ its closure.

Theorem 1 *Assume that Θ is convex, compact and $\Theta \subset \overline{B}(0,R)$ with $R \geq 0$. In addition, assume that $\theta \mapsto -\log(p(y|\theta)) + g(\theta) \in C^2(\Theta,\mathbb{R})$ and is convex, that $(x,\theta) \mapsto \log(p(x,y|\theta)) \in C^2(\mathbb{R}^d \times \Theta, \mathbb{R})$ and that there exists $m_1 > 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$*

$$\sup_{x\in\mathbb{R}^d} \|\nabla_\theta \log(p(x,y|\theta))\| \exp[-(m_1/4)\sqrt{1+\|x\|^2}] < +\infty.$$

Assume that there exist L_1, L_2, R_1, $c \geq 0$ and $m_2 > 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$ we have

$$\|\nabla_x^2 \log(p(x,y|\theta))\| \leq L_1,$$
$$\|\nabla_x \nabla_\theta \log(p(x,y|\theta))\| \leq L_2 \exp[(m_1/2)\sqrt{1+\|x\|^2}],$$
$$\langle \nabla_x \log(p(x,y|\theta)), x\rangle \leq -m_1\|x\|\mathbb{1}_{\mathrm{B}(0,R_1)^c}(x) - m_2\|\nabla_x \log(p(x,y|\theta))(x)\|^2 + c.$$

Let $(\gamma_n, \delta_n)_{n\in\mathbb{N}} \in ((\mathbb{R}_+^*)^2)^{\mathbb{N}}$ *non-increasing,* $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$ *non-decreasing and* $\sum_{n\in\mathbb{N}} \delta_n = +\infty$. *Assume*

$$\sum_{n\in\mathbb{N}} \delta_n(\gamma_n^{1/2} + (m_n\gamma_n)^{-1}) < +\infty\,, \tag{8}$$

or alternatively

$$\sum_{n\in\mathbb{N}} \delta_n(\gamma_n^{1/2} + \delta_n/\gamma_n^2 + (\gamma_{n+1} - \gamma_n)/\gamma_n^3) < +\infty\,. \tag{9}$$

Then almost surely, $\hat{\theta}_\infty = \lim_{N\to+\infty} \hat{\theta}_N$ *exists and* $\hat{\theta}_\infty \in \arg\min_\Theta f$ *for* δ_0, γ_0 *sufficiently small.*

Note that (8) only holds if $\lim m_n = +\infty$ (increasing batch size setting) whereas (9) holds even if $m_n = 1$ under tighter conditions on $(\delta_n)_{n\in\mathbb{N}}$ and $(\gamma_n)_{n\in\mathbb{N}}$ (fixed batch size setting). For constant sequences $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$, Theorem 1 does not apply and $(f(\hat{\theta}_N))_{N\in\mathbb{N}}$ is biased. However, we can control the asymptotic bias using the following result.

Theorem 2 *Under the same conditions on* Θ, $(x, \theta) \mapsto \log(p(x, y|\theta))$ *and* $\theta \mapsto -\log(p(y|\theta)) + g(\theta)$ *as in Theorem 1, there exist* $\bar{\delta}, \bar{\gamma} > 0$ *and* $C \geq 0$ *such that if for any* $n \in \mathbb{N}$, $\delta_n = \delta \in (0, \bar{\delta}]$, $\gamma_n = \gamma \in (0, \bar{\gamma}]$ *and* $m_n = 1$ *then* $\limsup_{N\to+\infty}\{\mathbb{E}[f(\hat{\theta}_N)] - \min_\theta f\} \leq C\gamma^{1/2}$.

We believe Theorem 2 to be highly relevant to practitioners, as we often empirically observe that the best trade-off between accuracy and computing-time is obtained by setting $(\gamma_n)_{n\in\mathbb{N}}$ to a constant and relatively large value (determined, e.g. from a bound on the Lipschitz constant of the target density). In our experience, this leads to a fast converging sequence that stabilises quickly close to the MLE, albeit with some bias.

The detailed theoretical analysis of the proposed SOUL method and its generalization is reported to Sect. 3. To conclude, Sect. 4 demonstrates the proposed methodology with three numerical experiments related to high-dimensional logistic regression and statistical audio analysis with sparsity promoting priors. Finally, we also study the case where f is not convex. In that case, we use the results of Kushner and Yin (2003) to establish that $(\theta_n)_{n\in\mathbb{N}}$ converges a.s to a stationary point of the projected ordinary differential equation associated with ∇f and Θ. We postpone this result to Section S3 in De Bortoli et al. (2019).

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

In this section, we state our main theoretical results in a broader framework than the one introduced in Sect. 2. After establishing notation and conventions in Sects. 3.1, 3.2 presents the general stochastic optimisation setting considered in this paper, which encompasses the MLE estimation problem (1). In order to address this class of optimisation problems, we develop a generalisation of SOUL: the Stochastic Optimisation via Unadjusted Kernel (SOUK) method. In this class of methods, ULA is replaced by a generic inexact Markov chain Monte Carlo method. We then establish our main results regarding the convergence properties of SOUK in Sect. 3.3. We conclude the section by showing that our general results apply to the specific MLE optimisation problem (1), and to the specific Langevin algorithm (6) used in SOUL in Sect. 3.4. All the proofs are postponed to the supplementary document.

3.1 Notation and convention

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$, and for $f : \mathbb{R}^d \to \mathbb{R}$ measurable, $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and f a μ-integrable function, denote by $\mu(f)$ the integral of f with respect to μ. For $f : \mathbb{R}^d \to \mathbb{R}$ measurable, 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

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

If $V = 1$, then $\|\cdot\|_V$ is the total variation denoted by $\|\cdot\|_{\mathrm{TV}}$. 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 and k-differentiable functions. 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 $m \geq 0$ if for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$,

$$f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)-\mathrm{m}t(1-t)\,\|x-y\|^2/2\,.$$

For any $a \in \mathbb{R}^d$ and $R > 0$, denote $\mathrm{B}(a, R)$ the open ball centered at a with radius R. Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A Markov kernel P is a mapping $\mathrm{K} : \mathsf{X}\times\mathcal{Y} \to [0, 1]$ such that for any $x \in \mathsf{X}$, $\mathrm{P}(x, \cdot)$ is a probability measure and for any $\mathsf{A} \in \mathcal{Y}$, $\mathrm{P}(\cdot, \mathsf{A})$ is measurable. For any probability measure μ on $(\mathsf{X}, \mathcal{X})$ and measurable function $f : \mathsf{Y} \to \mathbb{R}_+$ we denote $\mu\mathrm{P} = \int_{\mathsf{X}} \mathrm{P}(x, \cdot)\mathrm{d}\mu(x)$ and $\mathrm{P}f = \int_{\mathsf{Y}} f(y)\mathrm{P}(\cdot, \mathrm{d}y)$. In what follows the Dirac mass at $x \in \mathbb{R}^d$ is denoted by δ_x (which should not be confused with the stepsize sequence $(\delta_n)_{n\in\mathbb{N}}$). The complement of a set $\mathsf{A} \subset \mathbb{R}^d$, is denoted by A^{c}. All densities are w.r.t. the Lebesgue measure unless stated otherwise.

3.2 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 assumption.

A 1 *(i) Θ is convex, closed, $\Theta \subset \mathrm{B}(0, M_\Theta)$ for $M_\Theta > 0$.*
(ii) There exist an open set $U \subset \mathbb{R}^{d_\Theta}$ and $L_f \geq 0$ such that $\Theta \subset U$, $f \in C^1(U, \mathbb{R})$ and for any $\theta_1, \theta_2 \in \Theta$

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

(iii) 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

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

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 $\theta \in \Theta$, $H_\theta : x \mapsto -\nabla_\theta \log(p(x, y|\theta)) + \nabla g(\theta)$ and π_θ is the probability distribution with density $x \mapsto p(x|y, \theta)$.

To minimise the objective function f we suggest the use of a SA strategy which extends the one presented in Sect. 2. More precisely, motivated by the methodology described in Sect. 2, we propose a SA scheme which relies on biased estimates of $\nabla f(\theta)$ through a family of Markov kernels $\{\mathrm{K}_{\gamma,\theta}, \gamma \in (0, \bar{\gamma}], \ \theta \in \Theta\}$, 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 $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. $\bar{\gamma}$ is an extra parameter which ensures the stability of the Markov kernel. 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 $(\Omega, \mathcal{F}, \mathbb{P})$, $(X_{1:m_n}^n, \theta_n)_{n\in\mathbb{N}}$ by the following recursion for $n \in \mathbb{N}$ and $k \in \{0, \ldots, m_n - 1\}$

$$\begin{aligned} &X_{1:m_n}^n \text{ is Markov chain with kernel } \mathrm{K}_{\gamma_n,\theta_n} \\ &\quad \text{and } X_0^n = X_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1} \, , \\ &\theta_{n+1} = \Pi_\Theta[\theta_n - \tfrac{\delta_{n+1}}{m_n} \textstyle\sum_{k=1}^{m_n} H_{\theta_n}(X_k^n)] \, , \end{aligned} \tag{10}$$

where Π_Θ is the projection onto Θ and $\mathcal{F}_n$ is defined by (7). By (10), for any $n \in \mathbb{N}$, θ_{n+1} is $\mathcal{F}_n$-measurable. Then, the sequence of approximate minimisers of f is given by $(\hat{\theta}_N)_{N\in\mathbb{N}}$, see (4). The recursion (10) defines the SOUK methodology.

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 a.s 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 exist $C, \alpha > 0$ 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 (Kushner and Yin 2003) which establish that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s to a stationary point of the projected ordinary differential equation associated with ∇f and Θ. Note that we restrict ourselves to the study of the convergence of $(\theta_n)_{n\in\mathbb{N}}$ and do not derive non-asymptotic bounds. We postpone this result to Section S3 in De Bortoli et al. (2019), since it involves a theoretical background which we think is out of the scope of the main document.

The SOUK methodology allows for the use of Markov kernels beyond the one associated with the ULA algorithm considered in Sect. 2.4. Important examples include the Moreau Yosida ULA and Proximal ULA algorithms see De Bortoli et al. (2020a), De Bortoli et al. (2020b). We are currently investigating application of the SOUK framework using other samplers, in particular the Stochastic Gradient Langevin Dynamics (Welling and Teh 2011b), the underdamped Langevin algorithm (Ma et al. 2019) and the Hamiltonian Monte Carlo algorithm (Girolami and Calderhead 2011; Durmus et al. 2017; Maddison et al. 2018).

Finally note that this general optimisation setting encompasses many cases of interest which are generalizations of the SOUL algorithm. If $(\mathsf{W}, \mathcal{W}, \mu_\mathsf{W})$ is a probability space and $f(\theta) = \int_\mathsf{W} \hat{f}(\theta, w) \mathrm{d}\mu_\mathsf{W}(w)$ for any $\theta \in \Theta$, with $\hat{f}$ such that for any $w \in \mathsf{W}$, $\hat{f}(\cdot, w)$ satisfies A1 with $\pi_\theta \leftarrow \mathrm{P}_\theta(w, \cdot)$, $H_\theta \leftarrow H_\theta(\cdot, w)$, where for any $\theta \in \Theta$, $\mathrm{P}_\theta : \mathsf{W} \times \mathcal{X} \to [0, 1]$ is a Markov kernel. Assume that $\int_\mathsf{W} \int_{\mathbb{R}^d} \|H_\theta(x, w)\| \mathrm{P}_\theta(w, \mathrm{d}x) \mathrm{d}\mu_\mathsf{W}(w) < +\infty$ for any $\theta \in \Theta$. Then, we can consider the following algorithm

$$W_n \text{ is a sample from } \mu_\mathsf{W}$$

$X^n_{1:m_n}$ is Markov chain with kernel $\mathrm{K}_{\gamma_n,\theta_n,W_n}$
and $X^n_0 = X^{n-1}_{m_{n-1}}$ given $\mathcal{F}_{n-1}$,
$\theta_{n+1} = \Pi_\Theta[\theta_n - \frac{\delta_{n+1}}{m_n}\sum_{k=1}^{m_n} H_{\theta_n}(X^n_k, W_n)]$,

where $\{\mathrm{K}_{\gamma,\theta,w} : \theta \in \Theta,\ \gamma \in (0, \bar{\gamma}\rceil, w \in \mathsf{W}\}$ is a family of Markov kernels. Similar convergence guarantees and control of the bias as the ones established in Sect. 3.3 can be obtained for this algorithm which allows to tackle the case where f can be written as a sum of functions.

3.3 Main results

We impose a stability condition on the stochastic process $X^n_{1:m_n}$ defined by (10) and that for any $\gamma \in (0, \bar{\gamma}\rceil$ and $\theta \in \Theta$ the iterates of $\mathrm{K}_{\gamma,\theta}$ are close enough to π_θ after a sufficiently large number of iterations.

H 1 *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 $n, p \in \mathbb{N}$, $k \in \{0, \ldots, m_n\}$

$$\mathbb{E}_{1=}[K^p_{\gamma_n,\theta_n} V(X^n_k)|X^0_0] \leq A_1 V(X^0_0) \ , \ \mathbb{E}\left[V(X^0_0)\right] < +\infty \ .$$

(ii) For any $\gamma \in (0, \bar{\gamma}\rceil$, $\theta \in \Theta$, $K_{\gamma,\theta}$ admits a stationary distribution $\pi_{\gamma,\theta}$ and there exist $A_2, A_3 \geq 1$, $\rho \in [0, 1)$ such that for any $\gamma \in (0, \bar{\gamma}\rceil$, $\theta \in \Theta$, $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\| \delta_x K^n_{\gamma,\theta} - \pi_{\gamma,\theta} \|_{1=V} \leq A_2 \rho^{n\gamma} V(x) \ , \qquad \pi_{\gamma,\theta}(V) \leq A_3 \ .$$

(iii) There exists $\boldsymbol{\Psi} : \mathbb{R}^\star_+ \to \mathbb{R}_+$ such that for any $\gamma \in (0, \bar{\gamma}\rceil$ and $\theta \in \Theta$, $\|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \leq \boldsymbol{\Psi}(\gamma)$.

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 (Meyn and Tweedie 1992; Douc et al. 2018). Many MCMC algorithms enjoy geometric ergodicity such as the independence sampler (Tierney 1994), the Random Walk Metropolis–Hastings algorithm (Jarner and Hansen 2000), the Hamiltonian Monte-Carlo algorithm (Durmus et al. 2017) or the ULA algorithm (Dalalyan 2017; Durmus and Moulines 2017). However, obtaining bounds which are independent from $\theta \in \Theta$ can be arduous. In Sect. 3.4, we show that such bounds can be established for ULA under regularity and curvature conditions on the family of potentials $\{U_\theta : \theta \in \Theta\}$ if for any $\theta \in \Theta$, π_θ admits a density proportional to $x \mapsto \exp[-U_\theta(x)]$. A popular way to establish geometric ergodicity is to derive minorization and Foster-Lyapunov drift conditions (Hairer and Mattingly (2011); De Bortoli and Durmus (2019)) which can be verified on a case-by-case basis depending on the Markov kernel at hand, see Douc et al. (2018) for instance. H1-(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 S15 in De Bortoli et al. (2019). We now state our mains results.

Theorem 3 (Increasing batch size 1) *Assume H1 and that 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 $\delta_0 < 1/L_f$, $\gamma_0 < \bar{\gamma}$ and*

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \ , \ \sum_{n=0}^{+\infty} \delta_{n+1}(\boldsymbol{\Psi}(\gamma_n) + (m_n\gamma_n)^{-1}) < +\infty \ . \tag{11}$$

Let $(\theta_n)_{n\in\mathbb{N}}$ and $(X^n_{1:m_n})_{n\in\mathbb{N}}$ be given by (10). Assume in addition that H1 is satisfied and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. Then, the following statements hold:

(a) $(\theta_n)_{n\in\mathbb{N}}$ converges a.s to some $\theta^\star \in \arg\min_\Theta f$;
(b) furthermore, a.s there exists $C \geq 0$ such that for any $n \in \mathbb{N}^*$

$$f(\hat{\theta}_n) - \min_\Theta f \leq C \big/ \left(\textstyle\sum_{k=1}^n \delta_k\right) \ .$$

Proof The proof is postponed to Section S1.1 in De Bortoli et al. (2019). □

Note that in (10), $X^n_0 = X^{n-1}_{m_{n-1}}$ for $n \in \mathbb{N}^*$. This procedure is referred to as warm-start in the sequel. An inspection of the Proof of Theorem 3 shows that X^n_0 could be any random variable independent from $\mathcal{F}_{n-1}$ for any $n \in \mathbb{N}$ with $\sup_{n\in\mathbb{N}^*} \mathbb{E}\left[V(X^n_0)\right] < +\infty$. It is not an option in the fixed batch size setting of Theorem 5, where the warm-start procedure is crucial for the convergence to occur.

We extend this theorem to non-convex objective function, see Section S3 in De Bortoli et al. (2019). Under the conditions of Theorem 3 with the additional assumption that Θ is a smooth manifold we obtain that $(\theta_n)_{n\in\mathbb{N}}$ converges a.s to some point $\theta^\star$ such that $\nabla f(\theta^\star) + \mathbf{n} = 0$ with $\mathbf{n} = 0$ if $\theta^\star \in \mathrm{int}(\Theta)$ and $\mathbf{n} \in \mathrm{N}(\theta^\star, \Theta)$ where $\mathrm{N}(\theta^\star, \Theta)$ is the convex cone generated by the outer normals at point $\theta^\star$, see (Aubin 2000, 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 S15 in De Bortoli et al. (2019) shows that H1-(iii) holds with for any $\gamma \in (0, \bar{\gamma}\rceil$, $\boldsymbol{\Psi}(\gamma) = \mathcal{O}(\gamma^{1/2})$. Assume 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 (11) is equivalent

to

$$a \leq 1 \,, \qquad a + b/2 > 1 \,, \qquad a - b + c > 1 \,. \tag{12}$$

Suppose $a \in [0, 1]$ is given, (12) reads

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

This illustrates the trade-off between the intrinsic inaccuracy of our algorithm through the family of Markov kernels (10) 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. In this worst-case scenario, the convergence is guaranteed if $\gamma_n = n^{-2-\varsigma_1}$ 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 4 (Increasing batch size 2) *Assume A1 and that 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 $\delta_0 < 1/L_f$, $\gamma_0 < \bar{\gamma}$. Let $(\theta_n)_{n\in\mathbb{N}}$ and $(X^n_{1:m_n})_{n\in\mathbb{N}}$ be given by (10). Assume in addition that H1 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 $n \in \mathbb{N}^*$*

$$\mathbb{E}[f(\hat{\theta}_n) - \min_\Theta f] \leq E_n \big/ \left(\textstyle\sum_{k=1}^n \delta_k\right) \,,$$

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

$$\begin{aligned} E_n &= 2M_\Theta^2 + 2B_1 M_\Theta \mathbb{E}[V^{1/2}(X_0^0)] \sum_{k=0}^{n-1} \delta_{k+1}/(m_k\gamma_k) \\ &\quad + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1} \boldsymbol{\Psi}(\gamma_k) \\ &\quad + 4B_1^2 \mathbb{E}\left[V(X_0^0)\right] \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 \\ &\quad + 4 \sum_{k=0}^{n-1} \delta_{k+1}^2 \boldsymbol{\Psi}(\gamma_k)^2 + B_2 \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 \,, \end{aligned} \tag{13}$$

with B_1, B_2 given in Lemmas S3 and S4 in De Bortoli et al. (2019).

Proof The proof is postponed to Section S1.2 in De Bortoli et al. (2019). □

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 S15 in De Bortoli et al. (2019) shows that for any $\gamma \in (0, \bar{\gamma}]$, $\boldsymbol{\Psi}(\gamma) = \mathcal{O}(\gamma^{1/2})$. 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 (12) 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 $\varsigma_2 > 0$. Therefore, for a fixed target precision $\varepsilon > 0$, it requires that $\varepsilon = \mathcal{O}(n^{a-1})$ and the complexity reads $\mathcal{O}(\varepsilon^{-3}\,(\log(1/\varepsilon)/(1-a))^{1+\varsigma_2})$. 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 matches the order identified in the deterministic gradient descent for convex functionals, see (Bertsekas 1997, Proposition 1.3.3) for instance in the unconstrained case. This behavior is specific to the increasing batch size setting.

Another case of interest is the fixed 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 (Pólya and Szegő 1998, Problem 80, Part I), it holds that

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

Therefore, we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}[f(\hat{\theta}_N) - \min f] \leq 2M_\Theta \boldsymbol{\Psi}(\gamma) \,.$$

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

$$\limsup_{n\to+\infty} \mathbb{E}[f(\hat{\theta}_N) - \min f] \leq \boldsymbol{\Xi}_1(\gamma) \,, \tag{14}$$

with $\boldsymbol{\Xi}_1(\gamma) = 2B_1 M_\Theta \mathbb{E}[V^{1/2}(X_0^0)]/\gamma + 2M_\Theta \boldsymbol{\Psi}(\gamma)$. However if $(m_n)_{n\in\mathbb{N}}$ is constant the convergence cannot be obtained using Theorem 3. Strengthening the conditions of Theorem 3 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 (10) can be written for any $n \in \mathbb{N}$ as

$$\begin{aligned} &\tilde{X}_{n+1} \text{ has distribution } \mathrm{K}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n, \cdot) \text{ conditional to } \tilde{\mathcal{F}}_n \,, \\ &\tilde{\theta}_{n+1} = \Pi_\Theta \left[\tilde{\theta}_n - \delta_{n+1} H_{\tilde{\theta}_n}(\tilde{X}_{n+1}) \right] \,, \end{aligned} \tag{15}$$

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

$$\tilde{\mathcal{F}}_n = \sigma\left(\tilde{\theta}_0, (\tilde{X}_\ell)_{\ell\in\{0,\ldots,n\}}\right) \,. \tag{16}$$

We consider the following assumption.

A 2 *There exists* $L_H \geq 0$ *such that for any* $x \in \mathbb{R}^d$ *and* $\theta_1, \theta_2 \in \Theta$, $\|H_{\theta_1}(x) - H_{\theta_2}(x)\| \leq L_H \|\theta_1 - \theta_2\| V^{1/2}(x)$, *where* V *is given in H1.*

We consider a similar property as A2 on the family of Markov kernels $\{\mathrm{K}_{\gamma,\theta}, \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$, which weakens the assumption (Atchadé et al. 2017, H6). Indeed, we do not assume that for any $\gamma \in (0, \bar{\gamma}]$, $\theta \mapsto \mathrm{K}_{\gamma,\theta}$ is Lipschitz.

H 2 *There exist* $\boldsymbol{\Lambda}_1 : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ *and* $\boldsymbol{\Lambda}_2 : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ *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* $a \in [1/4, 1/2]$

$$\begin{aligned}&\| \underset{x}{\delta} K_{\gamma_1,\theta_1} - \underset{x}{\delta} K_{\gamma_2,\theta_2} \|_{V^a} \\ &\quad \leq \left[\boldsymbol{\Lambda}_1(\gamma_1, \gamma_2) + \boldsymbol{\Lambda}_2(\gamma_1, \gamma_2)\|\theta_1 - \theta_2\|\right] V^{2a}(x) \,,\end{aligned}$$

where V *is given in H1.*

The following theorem ensures convergence properties for $(\theta_n)_{n\in\mathbb{N}}$ similar to the ones of Theorem 3. The proof of this result is based on a generalization of (Fort et al. 2011, Lemma 4.2) for inexact MCMC schemes.

Theorem 5 (Fixed batch size 1) *Assume A1, A2 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* $\delta_0 < 1/L_f$, $\gamma_0 < \bar{\gamma}$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n|\delta_n^{-2} < +\infty$, $\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty$ *and*

$$\begin{aligned}&\sum_{n=0}^{+\infty} \delta_{n+1}\gamma_{n+1}^{-2} \left[\boldsymbol{\Lambda}_1(\gamma_n, \gamma_{n+1}) + \delta_{n+1}\boldsymbol{\Lambda}_2(\gamma_n, \gamma_{n+1})\right] < +\infty \,, \\ &\sum_{n=0}^{+\infty} \delta_{n+1}\boldsymbol{\Psi}(\gamma_n) < +\infty \,, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}^2\gamma_n^{-2} < +\infty \,. \qquad (17)\end{aligned}$$

Let $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ *and* $(\tilde{X}_n)_{n\in\mathbb{N}}$ *be given by (15). Assume in addition that H1 and H2 are satisfied and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x)$. *Then, the following statements hold:*

(a) $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ *converges a.s to some* $\theta^\star \in \arg\min_\Theta f$;
(b) furthermore, a.s there exists $C \geq 0$ *such that for any* $n \in \mathbb{N}^*$

$$f(\hat{\theta}_n) - \min_\Theta f \leq C \Big/ \left(\textstyle\sum_{k=1}^n \delta_k\right) \,.$$

Proof The proof is postponed to Section S1.3 in De Bortoli et al. (2019). □

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, Propositions S15 and S16 in De Bortoli et al. (2019) show that for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and $\gamma_1 > \gamma_2$, $\boldsymbol{\Psi}(\gamma_1) = \mathrm{C}_1\gamma_1^{1/2}$, $\boldsymbol{\Lambda}_1(\gamma_1, \gamma_2) = \mathrm{C}_2(\gamma_1/\gamma_2 - 1)$, $\boldsymbol{\Lambda}_2(\gamma_1, \gamma_2) = \mathrm{C}_3\gamma_2^{1/2}$, and $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 \geq 0$. Thus, we obtain that the following series should converge

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

In addition, assume that $\delta_n = n^{-a}$ and that $\gamma_n = n^{-b}$ with $a, b > 0$. In this case, the summability conditions of Theorem 5 read

$$a \leq 1 \,, \; a + b/2 > 1 \,, \; 2a - 2b > 1 \,, \; a + (b+1) - 3b > 1 \,,$$

i.e. $b \in \mathsf{I} = (2(1-a), a - 1/2)$ and $a \in [0, 1]$. Note that $\mathsf{I} \neq \emptyset$ as soon as $a > 5/6$. In the special setting where $a = 1$ then the convergence in Theorem 5 occurs if $b \in (0, 1/2)$. Since $a > b + 1/2$ we obtain that $\lim_{n\to+\infty}(\delta_n/\gamma_n) = 0$. This means that the stochastics gradient descent dynamic associated with $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ moves slower than the sequence $(\tilde{X}_n)_{n\in\mathbb{N}}$.

Theorem 6 (Fixed batch size 2) *Assume A1, A2 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* $\delta_0 < 1/L_f$ *and* $\gamma_0 < \bar{\gamma}$. *Let* $(\theta_n)_{n\in\mathbb{N}}$ *and* $(\tilde{X}_n)_{n\in\mathbb{N}}$ *be given by (15). Assume in addition that H1 and H2 are satisfied and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x)$. *Then, there exists* $(\tilde{E}_n)_{n\in\mathbb{N}}$ *such that for any* $n \in \mathbb{N}^*$

$$\mathbb{E}[f(\hat{\theta}_n) - \min_\Theta f] \leq \tilde{E}_n \Big/ \left(\textstyle\sum_{k=1}^n \delta_k\right) \,,$$

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

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

where C_1, C_2 *and* C_3 *are given in Lemmas S5, S8 and S7 in De Bortoli et al. (2019), respectively.*

Proof The proof is postponed to Section S1.4 in De Bortoli et al. (2019). □

Theorem 6 improves the conclusions of Theorem 4 in the case where $\gamma_n = \gamma > 0$ for any $n \in \mathbb{N}$. Indeed, in that case, similarly to (14), assuming that $\lim_{n\to+\infty} \delta_n = 0$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n| \, \delta_n^{-2} < +\infty$ and that for any $\gamma \in (0, \bar{\gamma}]$, $\boldsymbol{\Lambda}_1(\gamma, \gamma) = 0$ and we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}[f(\hat{\theta}_n) - \min f] \leq \boldsymbol{\Xi}_2(\gamma) \ ,$$

with

$$\begin{aligned} \boldsymbol{\Xi}_2(\gamma) &= 2M_\Theta \boldsymbol{\Psi}(\gamma) \\ &\leq \boldsymbol{\Xi}_1(\gamma) = 2B_1 M_\Theta \mathbb{E}[V^{1/2}(X_0^0)]/\gamma + 2M_\Theta \boldsymbol{\Psi}(\gamma) \ . \end{aligned}$$

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.

Note that the conclusions of Theorem 6 are similar to the ones of (Karimi et al. 2019, Theorem 2). In Karimi et al. (2019) the main result is a bound on $\mathbb{E}[\sum_{k=1}^n \delta_k \, \|\nabla_\theta f(\theta_k)\|^2 \, / \, \sum_{k=1}^n \delta_k]$ and $\nabla f(\theta)$ is not assumed to be convex but only related to a Lyapunov functional (Karimi et al. 2019, A1-A3). However, it is assumed that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ the invariant probability distribution of the Markov kernel $\mathrm{K}_{\gamma,\theta}$ is π_θ, *i.e.* $\boldsymbol{\Psi} = 0$ in H1-(iii), which is not the case in our analysis. Following this line of work, one could establish similar non-asymptotic result in the non-convex setting using the SOUK methodology. However, this is highly technical study that is beyond the scope of the present paper, and which we defer to future work.

To conclude, we highlight the differences between our work and Atchadé et al. (2017). Our results are based on the deterministic estimates derived in (Atchadé et al. 2017, Theorem 1, Theorem 2). However (Atchadé et al. 2017, Theorem 4, Theorem 6) rely on (i) the fact that for any $\theta \in \Theta$, π_θ is an invariant probability measure for $\mathrm{K}_{\gamma,\theta}$, see (Atchadé et al. 2017, H5) and (ii) a Lipschitz regularity property for $(\gamma, \theta) \mapsto \mathrm{K}_{\gamma,\theta}$, see (Atchadé et al. 2017, H6). Conditions (i) and (ii) do not hold if we consider *unadjusted* (inexact) Markov kernels. In this work we relax (Atchadé et al. 2017, H5) and (Atchadé et al. 2017, H6) by considering H1-(ii), respectively H2. As an important example, the results of Atchadé et al. (2017) do not apply if the Markov kernel is the one associated with ULA, whereas in Sect. 3.4 we show that our results do hold.

3.4 Application to SOUL

We now apply our results to the SOUL methodology introduced in Sect. 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). We consider the following assumption on the family of probability distributions $(\pi_\theta)_{\theta\in\Theta}$.

L 1 *For any $\theta \in \Theta$, there exists $U_\theta : \mathbb{R}^d \to \mathbb{R}$ such that π_θ admits a probability density function 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 $L \geq 0$ such that for any $x, y \in \mathbb{R}^d$,*

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

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 (10) can be rewritten for any $n \in \mathbb{N}$ and $k \in \{0, \ldots, m_n - 1\}$

$$\begin{aligned} &X_{k+1}^n = X_k^n - \gamma_n \nabla_x U_{\theta_n}(X_k^n) + \sqrt{2\gamma_n} Z_{k+1}^n \ , \\ &\text{with } X_0^n = X_{m_{n-1}}^{n-1} \text{ if } n \geq 1 \ , \\ &\theta_{n+1} = \Pi_\Theta[\theta_n - \tfrac{\delta_{n+1}}{m_n} \textstyle\sum_{k=1}^{m_n} H_{\theta_n}(X_k^n)] \ . \end{aligned} \tag{18}$$

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 also $(Z_k^n)_{n\in\mathbb{N},k\in\{1,\ldots,m_n\}}$ 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 above results hold by deriving sufficient conditions under which H1 and H2 are satisfied. 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 $\bar{\gamma} > 0$ which will be specified below.

L 2 *There exist $m_1, m_2 > 0$ and $c, R_1 \geq 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$,*

$$\langle \nabla_x U_\theta(x), x \rangle \geq m_1 \|x\| \mathbb{1}_{\mathrm{B}(0,R_1)^c}(x) + m_2 \|\nabla_x U_\theta(x)\|^2 - c \ .$$

L 3 *There exists $L_U \geq 0$ such that for any $x \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in \Theta$, $\left\|\nabla_x U_{\theta_1}(x) - \nabla_x U_{\theta_2}(x)\right\| \leq L_U \|\theta_1 - \theta_2\| V(x)^{1/2}$.*

The next theorems assert that under L1, L2 and L3, the SOUL algorithm introduced in Section 2 satisfies H1 and H2 and therefore Theorems 3, 4, 5 and 6 can be applied if in addition A1 and A2 hold. Under L2 define for any $x \in \mathbb{R}^d$

$$V_{\mathrm{e}}(x) = \exp[m_1 \sqrt{1 + \|x\|^2}/4] \ .$$

Theorem 7 *Assume L1 and L2. Then, H1 holds with $V \leftarrow V_e$, $\bar{\gamma} \leftarrow \min(1, 2m_2)$ and $\boldsymbol{\Psi}(\gamma) = D_4\sqrt{\gamma}$ where D_4 is given in Proposition S15 in De Bortoli et al. (2019).*

Proof The proof is postponed to Section S1.5 in De Bortoli et al. (2019). □

Theorem 8 *Assume L1, L2, L3 and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V_e^{1/4}(x)$. H2 holds with $V \leftarrow V_e$ and $\bar{\gamma} \leftarrow \min(1, 2m_2)$ and for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$, $\gamma_2 < \gamma_1$,*

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

where D_5 is given in Proposition S16 in De Bortoli et al. (2019).

Proof The proof is postponed to Section S1.6 in De Bortoli et al. (2019). □

Finally, we discuss the dependency of the complexity of the SOUL algorithm with respect to the dimension of the latent space d in the specific case where $\{U_\theta : \theta \in \Theta\}$ satisfies the following assumption.

L 4 *There exist $m_3 > 0$, $R_3 \geq 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$ with $\|x\| \geq R_3$, $\langle \nabla_x U_\theta(x), x \rangle \geq m_3 \|x\|^2$, U_θ is convex and there exist $C, \varpi_0 \geq 0$ such that*

$$\sup_{\theta \in \Theta} \{\|\nabla_x U_\theta(0)\| : \theta \in \Theta\} \leq C(1 + d^{\varpi_0}) \;,$$

where C and ϖ_0 are independent of the dimension d.

In what follows, we show that under L1, L3 and L4, the constants appearing in H1 and H2 depend polynomially on the dimension of the latent space d. This implies that the complexity of the SOUL algorithm is polynomial with respect to the dimension d since the constants appearing in Theorems 4 and 6 depend polynomially on the constants of H1 and H2.

Theorem 9 *Assume L1, L3 and L4. Then H1 and H2 are satisfied with $V : \mathbb{R}^d \to [1, +\infty)$ given for any $x \in \mathbb{R}^d$ by $V(x) = 1 + \|x\|^4$ and $\boldsymbol{\Psi}$, $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$ given by Theorems 7 and 8, respectively. Hence, the conclusions of Theorems 4 and 6 hold with $(E_n)_{n \in \mathbb{N}}$ and $(\tilde{E}_n)_{n \in \mathbb{N}}$ given for any $n \in \mathbb{N}$ by*

$$E_n = A(1 + d^{\varpi}) \left\{ \sum_{k=0}^{n} \delta_{k+1} \gamma_k^{1/2} + \sum_{k=0}^{n} \delta_{k+1}/(m_k \gamma_k) + \sum_{k=0}^{n} \delta_{k+1}^2/(m_k \gamma_k)^2 \right\} \;,$$

and

$$\tilde{E}_n = A(1 + d^{\varpi}) \left\{ \sum_{k=0}^{n} \delta_{k+1} \gamma_k^{1/2} + \sum_{k=0}^{n} |\delta_{k+1} - \delta_k| \, \gamma_k^{-1} + \sum_{k=0}^{n} \delta_{k+1} |\gamma_{k+1} - \gamma_k| \, \gamma_k^{-3} + \sum_{k=0}^{n} \delta_{k+1}^2 \gamma_k^{-1} + \delta_{n+1}/\gamma_n \right\} \;,$$

with $\varpi \in \mathbb{N}$ and $A \geq 0$ independent from d.

Proof The proof is postponed to Section S1.7 in De Bortoli et al. (2019) . □

4 Numerical results

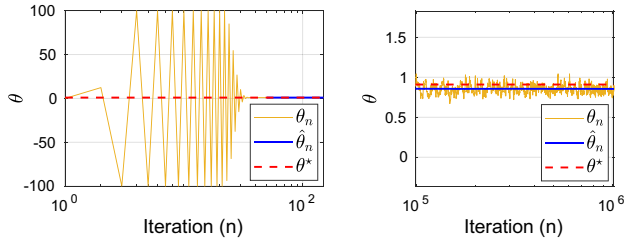
We now demonstrate the proposed methodology with three experiments that we have chosen to illustrate a variety of scenarios. Section 4.1 presents an application to empirical Bayesian logistic regression, where (1) can be analytically shown to be a convex optimisation problem with an unique solution $\theta^\star$, 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, Sect. 4.2 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 ($d = 2900$). Finally, Sect. 4.3 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@2.90 GHz workstation running MATLAB R2018a.

4.1 Bayesian Logistic Regression

In this first experiment we illustrate the proposed methodology with an empirical Bayesian logistic regression problem (Wakefield 2013; Polson et al. 2013). 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 $i \in \{1, \ldots, d_y\}$, 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, $\mathbf{1}_d = (1, \ldots, 1) \in \mathbb{R}^d$, $\sigma^2 = 5$ 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

$$p(y|\theta) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \left\{ \prod_{i=1}^{d_y} s(v_i^{\mathrm{T}}\beta)^{y_i} (1 - s(v_i^{\mathrm{T}}\beta))^{1-y_i} \right\} \times \exp[-\|\beta - \theta \mathbf{1}_d\|^2/(2\sigma^2)] \mathrm{d}\beta \;. \tag{19}$$

Lemma S18 in De Bortoli et al. (2019) shows that (19) is log-concave with respect to θ. In addition, using Lebesgue's dominated convergence theorem A1 is satisfied for any convex and compact set Θ with $H_\theta : \beta \mapsto -\nabla_\theta \log(p(\beta, y|\theta))$. We use the proposed SOUL methodology to estimate $\theta^\star$ for



(a) (b)

Fig. 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

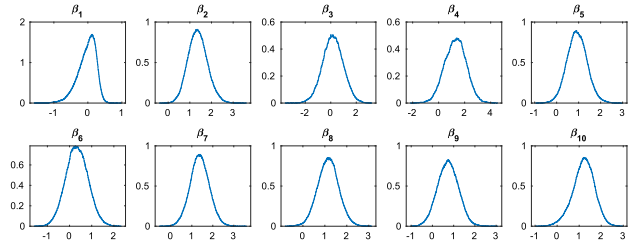


Fig. 2 Bayesian logistic regression-Normalised histograms of each component of β obtained with 2×10^6 Monte Carlo samples

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

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

Note that $\{\beta \mapsto -\log p(\beta|y, \theta) : \theta \in \Theta\}$ satisfies L1 and L2. Therefore, since A1 holds and $\theta \mapsto -\log p(y|\theta)$ is convex we get that Theorems 7 and 8 apply and the conclusions of Theorems 1 and 2 hold. 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 $\hat{\theta}_N$.

Figure 1a 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 1b presents the iterates θ_n for $n = 10^5, \ldots, 10^6$. For completeness, Fig. 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 true MLE $\theta^\star$ we use a truncated harmonic mean estimator (THME) Robert and Wraith (2009) 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. Given n samples $(\beta_i)_{i \in \{1,\ldots,n\}}$ from $p(\beta|y, \theta)$, we obtain an approximation of $p(y|\theta)$ by computing

$$\hat{p}(y|\theta) = n \, \mathrm{Vol}(\mathrm{B}(\bar{\beta}, R)) \Big/ \left(\textstyle\sum_{k=1}^{n} \frac{\mathbb{1}_{\mathsf{A}}(\beta_k)}{p(\beta_k, y|\theta)} \right) ,$$

with $\bar{\beta} = n^{-1} \sum_{k=1}^{n} \beta_k$, and radius $R \geq 0$ 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 Fig. 3a, 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 1000 times. Figure 3 shows the obtained histogram of $(\hat{\theta}_n)_{n \in \mathbb{N}}$, 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 discretisation 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

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

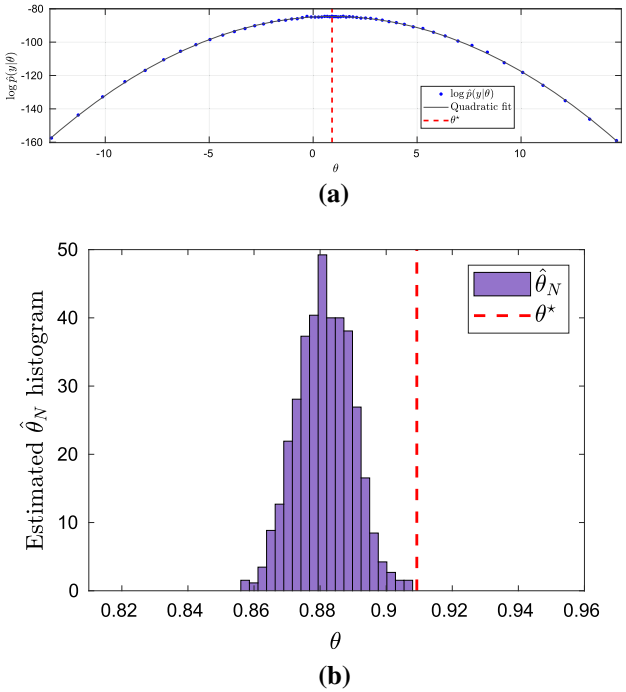


Fig. 3 Bayesian logistic regression-**a** Estimated points of the marginal log-likelihood $\log \hat{p}(y|\theta)$ with quadratic fit (corresponding to a Gaussian fit in linear scale). **b** Normalised histogram of $\hat{\theta}_N$ for 1000 repetitions of the experiment. An estimate of $\theta^\star$,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

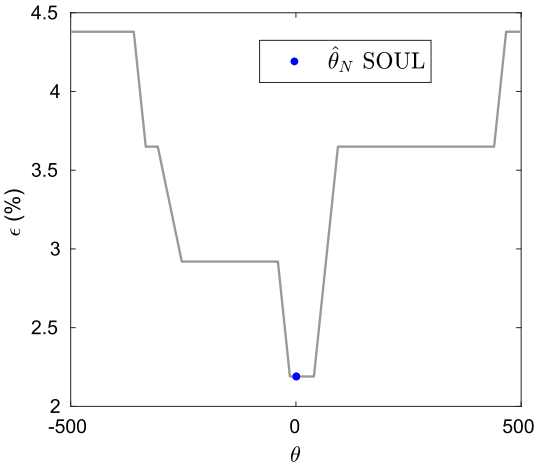


Fig. 4 Bayesian logistic regression-Percentage of mislabelled binary observations in terms of θ. In blue we show the value of $\hat{\theta}_N$ obtained with Algorithm 1

error $\epsilon = \sum_{i=1}^{d_{\text{test}}} \left| y_i^{\text{test}} - \hat{y}_i^{\text{test}} \right| / d_{\text{test}}$ and obtain $\epsilon = 2.2\%$. For comparison, Fig. 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.

4.2 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 (Candès et al. 2006; Candès and Wakin 2008). 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 $\boldsymbol{\Xi} \in \mathbb{R}^{\ell \times d}$, i.e, there exists a latent vector $x \in \mathbb{R}^d$ such that $z = \boldsymbol{\Xi} x$ and $\|x\|_0 = \sum_{i=1}^d \mathbb{1}_{\mathbb{R}^*}(x_i) \ll \ell$. This prior assumption is be modelled using a smoothed-Laplace distribution (Lingala and Jacob 2012)

$$p(x|\theta) \propto \exp\left[-(\theta/\lambda) \sum_{i=1}^d h_\lambda(x_i)\right] , \tag{20}$$

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

$$h_\lambda(u) = \begin{cases} u^2/2 & \text{if } |u| \le \lambda , \\ \lambda(|u| - \lambda/2) & \text{otherwise} . \end{cases} \tag{21}$$

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

$$p(y|x) \propto \exp\left[-\|y - \mathrm{A}x\|_2^2 / (2\sigma^2)\right] ,$$

where $\mathrm{A} = \mathbf{M}\boldsymbol{\Xi}$, leading to the posterior distribution

$$p(x|y) \propto \exp\left[-\|y - \mathrm{A}x\|_2^2 / (2\sigma^2) - (\theta/\lambda) \sum_{i=1}^d h_\lambda(x_i)\right] .$$

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

$$\hat{x}_{\text{MAP}} \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \|y - \mathrm{A}x\|_2^2 / 2\sigma^2 + (\theta/\lambda) \sum_{i=1}^d h_\lambda(x_i) \right\} , \tag{22}$$

and set $\hat{z}_{\text{MAP}} = \boldsymbol{\Xi} \hat{x}_{\text{MAP}}$.

Following decades of active research, there are now many convex optimisation algorithms that can be used to efficiently solve (22), even when d is very large (Chambolle and Pock 2016; Monga 2017). However, the selection of the value of θ in (22) 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 $\theta_{\text{cs}} = 0.1 \times \|\mathrm{A}^\top y\|_\infty / \sigma^2$, where for any $z \in \mathbb{R}^\ell$, $\|z\|_\infty = \max_{i \in \{1,\dots,\ell\}} |z_i|$, as suggested in Kim et al.

(2007) and Figueiredo et al. (2007); however, better results can 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 Balzano et al. (2010) 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 $d = 2900$ and is related to z by a dictionary matrix $\boldsymbol{\Xi}$ whose row vectors correspond to different piano notes lasting a quarter-second long [2]. The parameter λ for the prior (20) 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 (22) we use the Gradient Projection for Sparse Reconstruction (GPSR) algorithm proposed in Figueiredo et al. (2007). We use this solver because it is the one used in the online MATLAB demonstration of Balzano et al. (2010), 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.

Note that in this problem $\theta \mapsto -\log p(y|\theta)$ is non-convex whereas $x \mapsto -\log p(x|y,\theta)$ is convex. Using Lebesgue's dominated convergence theorem we get that A1 holds for any compact and convex set Θ. Note also that $\{x \mapsto -\log p(x|y,\theta) : \theta \in \Theta\}$ satisfies L1 and L2.

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 $\hat{\theta}_N$, Fig. 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 $\text{MSE}(\hat{x}_{\text{MAP}}) = \|z^\star - \boldsymbol{\Xi}\hat{x}_{\text{MAP}}\|_2^2/\ell$, 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.

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

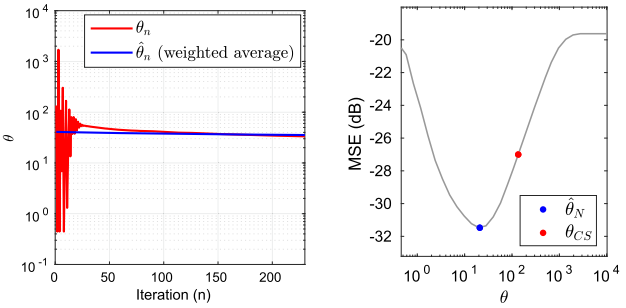


Fig. 5 Statistical audio compression-Evolution of the the iterate θ_n and $\hat{\theta}_n$ with $\sigma = 0.015$ in log scale (left). Reconstruction mean squared error (MSE) in dB as a function of the θ (right)

4.3 Sparse Bayesian logistic regression with random effects

Following on from the Bayesian logistic regression in Sect. 4.1, 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 is very large ($d_\theta = 1001$), making the MLE problem even more challenging. This experiment was previously considered by Atchadé et al. (2017) 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,

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

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 $\text{Ber}(\alpha)$ denotes the Bernoulli distribution with parameter $\alpha \in [0, 1]$ and $s(u) = \text{e}^u/(1 + \text{e}^u)$ 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 maximise (1), with marginal likelihood given by

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

and penalty function $g(\theta) = (\lambda\delta_0)^{-1}\sum_{j=1}^{d} h_{\lambda\delta_0}(\beta_j)$, where h_λ is the Huber function defined in (21). Using Lebesgue's dominated convergence theorem we get that A1 holds for any compact and convex set Θ.

We follow the procedure described in Atchadé et al. (2017) 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 is set to $\lambda = 30$. We emphasise 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 is $\log p(y|x,\theta) = \sum_{i=1}^{d_y}\{y_i(v_i^{\mathrm{T}}\beta + \sigma z_i^{\mathrm{T}}x) - \log(1+e^{v_i^{\mathrm{T}}\beta+\sigma z_i^{\mathrm{T}}x})\}$. To implement Algorithm 1 we use the gradients

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

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

Note that $\{x \mapsto -\log p(x|y,\theta) : \theta \in \Theta\}$ satisfies L1 and L2. Therefore, since A1 holds we get that Theorems 7 and 8 apply and the conclusions of Theorem S19 in De Bortoli et al. (2019) hold. Finally, the gradient of the penalty function is given by

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

where sign denotes the sign function, *i.e.* for any $s \in \mathbb{R}$, $\mathrm{sign}(s) = |s|/s$ if $s \neq 0$, and $\mathrm{sign}(s) = 0$ 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 (21) 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 $\tau = 0.005$).

[3] We renamed some symbols for notation consistency. What we denote by v_i, x, d_y and d, is denoted in Atchadé et al. (2017) by x_i, $\mathbf{U}$, N and q, respectively.

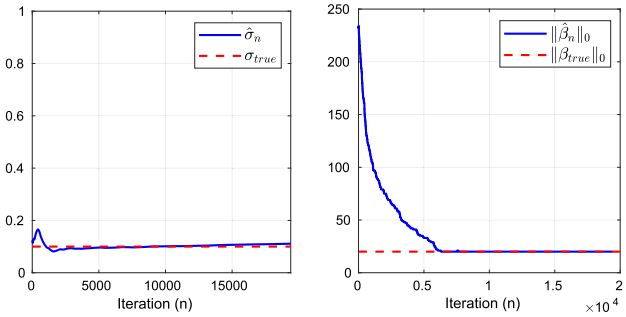


Fig. 6 Sparse Bayesian logistic regression with random effects-Evolution of the $\|\hat{\beta}_n\|_0$ and of the iterate $\hat{\sigma}_n$ for the proposed method. The true values are plotted in red as a reference

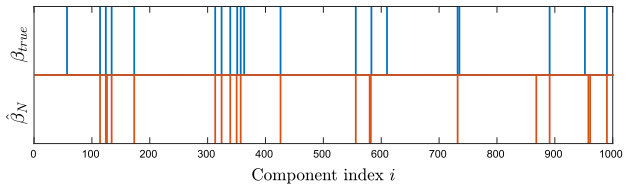


Fig. 7 Sparse Bayesian logistic regression with random effects-Support of the estimated $\hat{\beta}_N$ compared with the support of β_{true}

From Fig. 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, Fig. 7 shows that most active components were correctly identified. We also observe that $\hat{\beta}_n$ stabilises 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 (Atchadé et al. 2017, 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 Atchadé et al. (2017) considers the non-smooth penalty $g(\theta) = \lambda\|\beta\|_1$ instead of the Huber loss. Consequently, instead of using the gradient of g, they resort to the so-called proximal operator of g (Chambolle and Pock 2016). The generalisation of the SOUL methodology proposed in this paper to models that have non-differentiable terms is addressed in Vidal and Pereyra (2018), Vidal et al. (2019).

References

Ahn, S., Korattikara, A., Welling, M.: Bayesian posterior sampling via stochastic gradient fisher scoring. (2012) arXiv preprint arXiv:1206.6380,

Ahn, S., Shahbaba, B., Welling, M.: Distributed stochastic gradient mcmc. In: International Conference on Machine Learning, pp. 1044–1052, (2014)

Andrieu, C., Moulines, E.: On the ergodicity properties of some adaptive MCMC algorithms. Ann. Appl. Probab. **16**(3), 1462–1505 (2006)

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

Aubin, T.: A Course in Differential Geometry. Graduate Studies in Mathematics. AMS, New York (2000)

Balzano, L., Nowak, R., Ellenberg, J.: Compressed sensing audio demonstration. (2010) website http://web.eecs.umich.edu/~girasole/csaudio

Benveniste, A., Métivier, M., Priouret, P.: Adaptive algorithms and stochastic approximations, volume 22 of Applications of Mathematics (New York). Springer-Verlag, Berlin, (1990). Translated from the French by Stephen S. Wilson

Berger, R., Casella, G.: Statistical inference, 2nd edn. Duxbury / Thomson Learning, Pacific Grove, USA (2002)

Bertsekas, D.P.: Nonlinear programming. Journal of the Operational Research Society **48**(3), 334–334 (1997)

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

Candès, E.J. et al.: Compressive sampling. In: Proceedings of the international congress of mathematicians, volume 3, pages 1433–1452. Madrid, Spain, (2006)

Candès, E.J., Wakin, M.B.: 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)

Carlin, B.P., Louis, T.A.: Empirical Bayes: past, present and future. J. Am. Statist. Assoc. **95**(452), 1286–1289 (2000)

Casella, G.: An introduction to empirical Bayes data analysis. Am. Statist. **39**(2), 83–87 (1985)

Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numerica **25**, 161–319 (2016)

Cheng, X., Bartlett, P.: Convergence of langevin mcmc in kl-divergence. (2017). arXiv preprint arXiv:1705.09048

Cheng, X., Chatterji, N.S., Abbasi-Yadkori, Y., Bartlett, P.L., Jordan, M.I.: Sharp convergence rates for langevin dynamics in the nonconvex setting. (2018). arXiv preprint arXiv:1805.01648

Dalalyan, A.S.: Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. (2017). arXiv preprint arXiv:1704.04752

Dalalyan, A.S.: Theoretical guarantees for approximate sampling from smooth and log-concave densities. J. R. Stat. Soc. Ser. B. Stat. Methodol. **79**(3), 651–676 (2017)

Dalalyan, A.S., Riou-Durand, L.: On sampling from a log-concave density using kinetic langevin diffusions. (2018). arXiv preprint arXiv:1807.09382

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

De Bortoli, V., Durmus, A., Pereyra, M., Fernandez Vida, A.: Supplement to: Efficient stochastic optimisation by unadjusted langevin monte carlo. application to maximum marginal likelihood and empirical bayesian estimation. (2019)

De Bortoli, V., Durmus, A., Pereyra, M., Vidal, A.F.: Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach part ii: Theoretical analysis. SIAM J. Imaging Sci. **13**(4):1990–2028 (2020a). https://doi.org/10.1137/20M1339842

De Bortoli, V., Durmus, A., Vidal, A.F., Pereyra, M.: Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach. part ii: Theoretical analysis. arXiv preprint arXiv:2008.05793, (2020b)

Delyon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the em algorithm. Ann. Statist. **27**(1), 94–128 (1999)

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

Douc, R., Moulines, E., Priouret, P., Soulier, P.: Markov Chains. Springer, Berlin (2018). to be published

Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)

Durmus, A., Moulines, E.: Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. Ann. Appl. Probab. **27**(3), 1551–1587 (2017)

Durmus, A., Moulines, E., Pereyra, M.: Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. SIAM J. Imaging Sci. **11**(1), 473–506 (2018)

Durmus, A., Moulines, E., Saksman, E.: On the convergence of hamiltonian monte carlo. arXiv preprint arXiv:1705.00166 (2017)

Eberle, A.: Reflection couplings and contraction rates for diffusions. Probab. Theory Related Fields **166**(3–4), 851–886 (2016)

Eberle, A., Guillin, A., Zimmer, R.: Couplings and quantitative contraction rates for langevin dynamics. arXiv preprint arXiv:1703.01617 (2017)

Eberle, A., Majka, M.B.: Quantitative contraction rates for markov chains on general state spaces. arXiv preprint arXiv:1808.07033 (2018)

Figueiredo, M.A., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE J. Selected Topics Signal Process. **1**(4), 586–597 (2007)

Fort, G., Moulines, E.: Convergence of the Monte Carlo expectation maximization for curved exponential families. Ann. Statist. **31**(4), 1220–1259 (2003)

Fort, G., Moulines, E., Priouret, P.: Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. Ann. Statist. **39**(6), 3262–3289 (2011)

Gentle, J.E., Härdle, W.K., Mori, Y.: Handbook of Computational Statistics: Concepts and Methods. Springer Science & Business Media, Berlin (2012)

Girolami, M., Calderhead, B.: Riemann manifold langevin and hamiltonian monte carlo methods. J. Royal Stat. Soc.: Ser. B (Stat. Methodol.) **73**(2), 123–214 (2011)

Hairer, M., Mattingly, J.C.: Yet another look at harris' ergodic theorem for markov chains. In: Seminar on Stochastic Analysis, Random Fields and Applications Vol. 63. pp. 109–117. Birkhäuser/Springer Basel AG, Basel (2011). https://doi.org/10.1007/978-3-0348-0021-1_7

Jarner, S.F., Hansen, E.: Geometric ergodicity of metropolis algorithms. Stoch. Process. Their Appl. **85**(2), 341–361 (2000)

Kallenberg, O.: Foundations of Modern Probability. Springer Science & Business Media, Berlin (2006)

Karimi, B., Miasojedow, B., Moulines, É., Wai, H.-T.: Non-asymptotic analysis of biased stochastic approximation scheme. arXiv preprint arXiv:1902.00629 (2019)

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: A method for large-scale l1-regularized least squares. IEEE J. Selected Topics Signal Process. **1**(4), 606–617 (2007)

Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications. volume 35 of Applications of

Mathematics (New York): Stochastic Modelling and Applied Probability, 2nd edn. Springer-Verlag, New York (2003)
Lee, H., Risteski, A., Ge, R.: Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In:Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.) Advances in Neural Information Processing Systems, Curran Associates, Inc. Vol. 31. pp. 7847–7856 (2018). https://proceedings.neurips.cc/paper/2018/file/c6ede20e6f597abf4b3f6bb30cee16c7-Paper.pdf
Lingala, S.G., Jacob, M.: A blind compressive sensing frame work for accelerated dynamic mri. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1060–1063. IEEE (2012)
Ma, Y.-A., Chatterji, N., Cheng, X., Flammarion, N., Bartlett, P., Jordan, M. I.: Is there an analog of nesterov acceleration for mcmc? arXiv preprint arXiv:1902.00996 (2019)
Maddison, C.J., Paulin, D., Teh, Y. W., O'Donoghue, B., Doucet, A.: Hamiltonian Descent Methods. arXiv preprint arXiv:1809.05042 (2018)
Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes. I. Criteria for discrete-time chains. Adv. in Appl. Probab. **24**(3), 542–574 (1992)
Monga, V.: Handbook of Convex Optimization Methods in Imaging Science. Springer, Berlin (2017)
Muehlebach, M., Jordan, M. I.: A dynamical systems perspective on nesterov acceleration. arXiv preprint arXiv:1905.07436 (2019)
Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2008)
Patterson, S., Teh, Y. W.: Stochastic gradient riemannian langevin dynamics on the probability simplex. In: Advances in neural information processing systems, pp. 3102–3110 (2013)
Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using pólya-gamma latent variables. J. Am. Stat. Assoc. **108**(504), 1339–1349 (2013)
Pólya, G., Szegő, G.: Problems and theorems in analysis. I. Classics in Mathematics. Springer-Verlag, Berlin, (1998). Series, integral calculus, theory of functions, Translated from the German by Dorothee Aeppli, Reprint of the 1978 English translation
Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer-Verlag, New York (2004)
Robert C. P., Wraith, D.: Computational methods for bayesian model choice. In: Aip Conference Proceedings, vol. 1193, pp. 251–262. AIP (2009)
Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli **2**(4), 341–363 (1996)
Teh, Y.W., Thiery, A.H., Vollmer, S.J.: Consistency and fluctuations for stochastic gradient langevin dynamics. J. Mach. Learn. Res. **17**(1), 193–225 (2016)
Tierney, L. Markov chains for exploring posterior distributions. Ann. Statist. 22(4):1701–1762 (1994). https://doi.org/10.1214/aos/1176325750
Vidal, A.F., Bortoli, V. D., Pereyra, M., Durmus, A.: Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach. part i: Methodology and experiments (2019)
Vidal, A.F., Pereyra, M.: Maximum likelihood estimation of regularisation parameters. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1742–1746. IEEE (2018)
Vollmer, S.J., Zygalakis, K.C., Teh, Y.W.: Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. J. Mach. Learn. Res. **17**(1), 5504–5548 (2016)
Wakefield, J.: Bayesian and Frequentist Regression Methods. Springer Science & Business Media, Berlin (2013)
Welling, M., Teh, Y. W.: Bayesian learning via stochastic gradient langevin dynamics. In: Proceedings of the 28th international Conference on Machine Learning (ICML-11), pp. 681–688 (2011a)

Welling, M., Teh, Y. W.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the International Conference on Machine Learning, pp. 681–688 (2011b)