Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach

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

[1]School of Mathematical and Computer Sciences, Heriot Watt University & Maxwell Institute for Mathematical Sciences.
[2]CMLA - École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, 94235 Cachan, France.

November 27, 2019

Abstract

Many imaging problems require solving an inverse problem that is ill-conditioned or ill-posed. Imaging methods typically address this difficulty by regularising the estimation problem to make it well- posed. This often requires setting the value of the so-called regularisation parameters that control the amount of regularisation enforced. These parameters are notoriously difficult to set a priori, and can have a dramatic impact on the recovered estimates. In this paper, we propose a general empirical Bayesian method for setting regularisation parameters in imaging problems that are convex w.r.t. the unknown image. Our method calibrates regularisation parameters directly from the observed data by maximum marginal likelihood estimation, and can simultaneously estimate multiple regularisation parameters. A main novelty is that this maximum marginal likelihood estimation problem is efficiently solved by using a stochastic proximal gradient algorithm that is driven by two proximal Markov chain Monte Carlo samplers, thus intimately combining modern high-dimensional optimisation and stochastic sampling techniques. Furthermore, the proposed algorithm uses the same basic operators as proximal optimisation algorithms, namely gradient and proximal operators, and it is therefore straightforward to apply to problems that are currently solved by using proximal optimisation techniques. We also present a detailed theoretical analysis of the proposed methodology, including asymptotic and non-asymptotic convergence results with easily verifiable conditions, and explicit bounds on the convergence rates. The proposed methodology is demonstrated with a range of experiments and comparisons with alternative approaches from the literature. The considered experiments include image denoising, non-blind image deconvolution, and hyperspectral unmixing, using synthesis and analysis priors involving the ℓ_1, total-variation, total-variation and ℓ_1, and total-generalised-variation pseudo-norms.

1 Introduction

Image estimation problems are ubiquitous in science and industry, and a central topic of research in imaging sciences. Canonical examples include, for instance, image denoising [52], image deblurring [22, 48], compressive sensing [28, 71], super-resolution [59, 87], tomographic reconstruction [23], image inpainting [38, 78], source separation [14, 13], fusion [79, 53], and phase retrieval [17, 45]. Solving these problems has stimulated significant advances in imaging methods, models, theory, and algorithms [46, 67, 58, 21].

*Email: af69@hw.ac.uk
†Email: valentin.debortoli@cmla.ens-cachan.fr
‡Email: m.pereyra@hw.ac.uk
§Email: alain.durmus@cmla.ens-cachan.fr
Part of this work has been presented at the 25th IEEE International Conference on Image Processing (ICIP) [85]

Most image estimation problems are ill-conditioned or ill-posed [46], a difficulty that imaging methods typically address by regularising the estimation problem to make it well posed. This can be achieved in different ways. For example, in the variational framework, regularization is introduced by using penalty functions that favour solutions with desired structural or regularity properties (e.g., smoothness, piecewise-regularity, sparsity, or constraints), see [21]. In the Bayesian statistical framework, regularisation arises from the use of informative prior distributions that also allow promoting solutions with expected structural or regularity properties [46]. Moreover, regularisation can be explicitly specified, or learnt from data using modern machine learning techniques. We refer the reader to [3] for an excellent introduction to variational, statistical, and machine learning regularisation approaches.

A main difficulty that arises when using any regularisation technique is deciding how much regularisation is appropriate, as different imaging modalities, instrumental setups, scenes, and noise conditions often require using very different amounts of regularisation. The amount of regularisation is usually explicitly controlled by some of the parameters of the model. The difficulty resides in that setting the value of these regularisation parameters a priori is notoriously difficult, particularly in problems that are ill-posed or ill-conditioned where the regularisation has a dramatic impact on the estimated solutions (see [57, 64, 27] and the illustrative example in Figure 1). As a result, there is significant interest in methods for setting regularisation parameters in an automatic, robust, and adaptive way.

Indeed, the developments of methods to automatically set regularisation parameters is a long-standing research topic in imaging sciences. Some methods such as generalized cross-validation [40], the L-curve [51, 42], the discrepancy principle [60, 12], and residual whiteness measures [2] operate by analysing the residual between the observed data and a prediction derived from the observation model. Such methods can perform well in certain imaging problems, but they are mainly limited to cases involving a single scalar regularisation parameter. Alternatively, methods based on Stein's unbiased risk estimator (SURE) have also received a lot of attention in the late [39, 35, 27]. These methods seek to select the value of the regularisation parameters by minimising SURE-based surrogates of the estimation mean squared error [35, 68, 39]. SURE methods can perform remarkably well in mildly ill-posed or ill-conditioned problems, but they generally struggle with problems that are more severely ill-conditioned or ill-posed [55]. Some recent works also consider learning regularisation parameters from a training dataset of clean images [82].

Lastly, the Bayesian statistical framework provides two main strategies for addressing unknown regularisation parameters: the hierarchical and the empirical [74, 57]. So far, imaging methods have mainly adopted the hierarchical strategy, where the unknown regularisation parameters are incorporated into the model to define an augmented posterior, and subsequently removed from the model by marginalisation or estimated jointly with the unknown image [67, 64]. This is the strategy that is adopted by most Markov chain Monte Carlo and variational Bayesian approaches reported in the literature (see e.g., [65, 8]).

In this work we propose to adopt an empirical Bayesian approach to estimate the regularisation parameters directly from the observed data in a fully automatic and unsupervised way. We focus on imaging problems that are convex w.r.t. the unknown image and that can be efficiently solved by using modern convex optimisation techniques once the regularisation parameters have been set [58, 21]. In a manner akin to [57, 4], we set regularisation parameters directly from the observed data by maximum marginal likelihood estimation. A main novelty of our work is that this maximum marginal likelihood estimation problem is efficiently solved by using a stochastic proximal gradient algorithm that is powered by two proximal Markov chain Monte Carlo samplers, thus intimately combining the strengths of modern optimisation and sampling techniques. In addition to being highly efficient and delivering remarkably accurate solutions, the proposed method can be readily implemented with the same tools that are used to construct optimisation algorithms to estimate the unknown image by maximum-a-posteriori estimation, namely proximal and gradient operators.

The remainder of the paper is organised as follows. Section 2 defines the class of imaging problems we consider and introduces basic necessary concepts of Bayesian inference. Section 3 presents the proposed empirical Bayesian method to calibrate regularisation parameters and discusses connections with hierarchical Bayesian approaches. Section 4 presents a detailed analysis of the theoretical properties of the proposed methodology, including easily verifiable conditions for convergence and quantitative convergence rates. Section 5 demonstrates the proposed methodology with a variety of non-blind image deblurring and imaging denoising problems involving scalar-valued regularisation parameters, followed by two challenging experiments that require setting vector-valued regularisation parameters, namely sparse hyperspectral image unmixing with the SUnSAL model [44], and image denoising using a To-

tal Generalised Variation regulariser [16], where the parameters have strong dependencies making the estimation problem particularly difficult. We report comparisons with several alternative approaches from the literature, including the discrepancy principle [60], the SURE-based SUGAR method [27], and the hierarchical Bayesian method described in [64]. Conclusions and perspectives for future work are finally reported in Section 6. Implementation guidelines and proofs are postponed to the appendix.

2 Problem Statement

Let $d, d_y, d_\Theta \in \mathbb{N}$ and let $\Theta \subset (0, +\infty)^{d_\Theta}$ be a convex compact set. We consider the estimation of an unknown image $x \in \mathbb{R}^d$ from an observation $y \in \mathbb{C}^{d_y}$ related to x by a statistical model with likelihood function

$$p(y|x) \propto \mathrm{e}^{-f_y(x)} ,$$

where f_y is convex and continuously differentiable with $\mathtt{L}_y$-Lipschitz gradient, *i.e.* for any $u, v \in \mathbb{R}^d$, $\|\nabla f_y(u) - \nabla f_y(v)\| \leqslant \mathtt{L}_y \|u - v\|$ where $\mathtt{L}_y > 0$. This class includes important observation models, in particular Gaussian linear models of the form $y = \mathrm{A}x + \mathrm{w}$ where $\mathrm{A} \in \mathbb{C}^{d_y \times d}$ and w is a d_y-dimensional Gaussian random variable with zero mean and covariance matrix $\sigma^2 \mathrm{Id}$ with $\sigma > 0$. We adopt a Bayesian approach and seek to use prior knowledge about x to regularise the estimation problem and improve results. We consider prior distributions given for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$ by

$$p(x|\theta) = \mathrm{e}^{-\theta^{\mathrm{T}} g(x)} / \mathrm{Z}(\theta) ,$$

for some convex and Lipschitz continuous vector of statistics $g : \mathbb{R}^d \to \mathbb{R}^{d_\Theta}$ and where we recall that the normalising constant of the prior distribution $p(x|\theta)$ is given by

$$\mathrm{Z}(\theta) = \int_{\mathbb{R}^d} \mathrm{e}^{-\theta^{\mathrm{T}} g(\tilde{x})} \mathrm{d}\tilde{x} . \tag{1}$$

Note that θ controls the amount of regularity enforced. The function g is allowed to be non-differentiable in order to include popular models such as $g(x) = \|\mathrm{B}x\|_\dagger$ for some dictionary $\mathrm{B} \in \mathbb{R}^{d_1 \times d}$ with $d_1 \in \mathbb{N}$ and norm $\|\cdot\|_\dagger$, as well as constraints on the solution space such as pixel-positivity.

Although rarely mentioned in the literature, these widely used prior distributions regularise the estimation problem by promoting solutions for which $g(x)$ is close to the expected value $\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|\theta) \, \mathrm{d}\tilde{x}$, which depends on θ. Formally, by differentiating (1) and using Leibniz integral rule [66] we obtain that for any $\theta \in \Theta$

$$\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|\theta) \, \mathrm{d}\tilde{x} = -\nabla_\theta \log \mathrm{Z}(\theta). \tag{2}$$

Additionally, because the prior distribution $x \mapsto p(x|\theta)$ is log-concave, using [15, Theorem 1.2] we have that for any $\varepsilon \in [0, 2]$

$$\int_{C_{\theta,\varepsilon}} p(\tilde{x}|\theta) \mathrm{d}\tilde{x} \leqslant 3 \exp\left[-\varepsilon^2 d/16\right] ,$$

with $C_{\theta,\varepsilon} = \{\tilde{x} \in \mathbb{R}^d \, : \, d^{-1} |\theta^{\mathrm{T}}(g(\tilde{x}) - \bar{g}_\theta)| \geqslant \varepsilon\}$. This result establishes that the prior distribution $x \mapsto p(x|\theta)$ strongly promotes solutions for which $g(x) \approx -\nabla_\theta \log \mathrm{Z}(\theta)$ with high probability when d is large.

Once the likelihood and prior $p(y|x)$ and $p(x|\theta)$ are specified, we use Bayes' theorem [74] to derive the posterior for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$

$$p(x|y, \theta) = p(y|x)p(x|\theta)/p(y|\theta) = \exp[-f_y(x) - \theta^{\mathrm{T}} g(x)] \Big/ \int_{\mathbb{R}^d} \exp[-f_y(\tilde{x}) - \theta^{\mathrm{T}} g(\tilde{x})] \mathrm{d}\tilde{x} \ . \tag{3}$$

This posterior distribution underpins all inferences about the image x given observed data y. In particular, imaging methods typically use the maximum-a-posteriori (MAP) estimator, given for any $\theta \in \Theta$ by

$$\hat{x}_{\theta,\mathrm{MAP}} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \{f_y(\tilde{x}) + \theta^{\mathrm{T}} g(\tilde{x})\} . \tag{4}$$

This Bayesian estimator has a number of favourable theoretical and computational properties (see [63] for a recent theoretical analysis of this estimator). From a computation viewpoint, since the posterior

$x \mapsto p(x|y,\theta)$ is log-concave, the computation of $\hat{x}_{\theta,\mathrm{MAP}}$ is a convex optimisation problem that can usually be efficiently solved using modern optimisation algorithms, see [21]. Imaging MAP algorithms typically adopt a proximal splitting approach [24] involving the gradient ∇f_y and the proximal operator of g, $\mathrm{prox}_g^\lambda : \mathbb{R}^d \to \mathbb{R}^{d_\Theta}$, see [11, Definition 12.23]. This operator is defined for any $\lambda > 0$ and $x \in \mathbb{R}^d$ by

$$\mathrm{prox}_g^\lambda(x) = \underset{\tilde{x}\in\mathbb{R}^d}{\mathrm{argmin}} \left\{ g(\tilde{x}) + \|\tilde{x} - x\|_2^2 / (2\lambda) \right\} ,$$

The smoothness parameter $\lambda > 0$ controls the regularity properties of the proximal operator. As mentioned previously, the regularization parameter $\theta \in \Theta$ controls the balance between observed and prior information, and can significantly impact inferences about the unknown image $x \in \mathbb{R}^d$, especially in problems that are ill-posed or ill-conditioned. In Figure 1, we illustrate the dramatic effect that the value of $\theta \in \Theta$ may have on the recovered image for a deconvolution problem with a total-variation prior. As expected, when θ is too small the estimated image is very noisy due to lack of regularisation, and when θ is too large the resulting image is over-regularised.

Figure 1: Deblurring of the `boat` image with total-variation prior (SNR=40 dB). Maximum-a-posteriori estimators for different values of $\theta > 0$ illustrating the effect of regularisation (increasing from left to right).

3 Proposed Empirical Bayes methodology

3.1 Empirical Bayes estimation

Under an empirical Bayesian paradigm, the regularisation parameter $\theta \in \Theta$ is estimated directly from the observed data y, for example by maximum marginal likelihood estimation. That is, we compute

$$\theta_\star \in \underset{\theta\in\Theta}{\mathrm{argmax}}\ p(y|\theta) \,, \tag{5}$$

where we recall that the marginal likelihood $p(y|\theta)$ is given for any $\theta \in \Theta$ by

$$p(y|\theta) = \int_{\mathbb{R}^d} p(y|\tilde{x}) p(\tilde{x}|\theta) \mathrm{d}\tilde{x} \,. \tag{6}$$

Given $\theta_\star$, empirical Bayesian approaches base inferences on the pseudo-posterior $x \mapsto p(x|y,\theta_\star)$, [18], given for any $x \in \mathbb{R}^d$ by

$$p(x|y,\theta_\star) = \exp[-f_y(x) - \theta_\star^\top g(x)] \Big/ \int_{\mathbb{R}^d} \exp[-f_y(\tilde{x}) - \theta_\star^\top g(\tilde{x})] \mathrm{d}\tilde{x} \ . \tag{7}$$

Observe that this strategy is equivalent to Bayesian model selection on a continuous class of models parametrised by θ, where $\theta_\star$ produces the model with the best fit-to-data (under some additional assumptions, $p(y|\theta_\star)$ provides the best approximation of the true distribution of y in a Kullback–Leibler divergence sense [86]).

Empirical Bayesian approaches were first considered in the statistical methodology community (see e.g. [72, 18]), which stimulated developments in computational statistics [73, 5, 6] to enable empirical Bayesian inference for general statistical models. This was recently followed by important theoretical works on the validity of the empirical approach and connections to the hierarchical Bayesian paradigm (see e.g. [69, 49, 76]).

Unfortunately, this powerful inference strategy is difficult to apply in imaging problems [77] because the marginal likelihood $\theta \mapsto p(y|\theta)$ is computationally intractable as it involves two d-dimensional integrals, namely (1) and (6), thus making the optimisation problem (5) very challenging. The aim of this paper is to enable empirical Bayesian inference in imaging inverse problems, with a focus on automatic selection of regularisation parameters for convex problems that would be typically solved by using proximal optimisation techniques. More precisely, inspired by [5, 6], we propose a stochastic gradient Markov chain Monte Carlo (MCMC) algorithm to efficiently solve (5) for imaging models of the general form (3), where two main novelties are that we use state-of-the-art proximal MCMC methods [33] to construct a stochastic optimisation scheme that scales efficiently to high dimensions, and that we provide easily verifiable theoretical conditions ensuring convergence.

Lastly, we note that the maximum likelihood estimation problem (5) raises natural questions about the uniqueness of $\theta_\star$, and about the log-concavity of the marginal likelihood $\theta \mapsto p(y|\theta)$, which are important for the convergence of iterative algorithms to compute $\theta_\star$. In particular, $p(y|\theta)$ could potentially admit more than one maximiser. However, we have not observed this in practice in any imaging problem. Indeed, because in our experiments $d_y \gg d_\Theta$, we suspect that the marginal likelihood $\theta \mapsto p(y|\theta)$ concentrates sharply around a single maximiser $\theta_\star$, and is strongly log-concave w.r.t. θ in the neighbourhood of $\theta_\star$. These favourable properties can be formally derived under simplifying assumptions (e.g. that $p(y|\theta)$ is fully separable on y [83]). Extending conditions for uniqueness of (5) to more general imaging problems is an important perspective for future work.

3.2 Stochastic gradient MCMC algorithm

We now present the proposed empirical Bayesian method to solve the marginal maximum likelihood estimation problem (5) and set regularisation parameters. As mentioned previously, the main difficulty in solving (5) is that the marginal likelihood function $\theta \mapsto p(y|\theta)$ is computationally intractable.

Suppose for now that $\theta \mapsto p(y|\theta)$ was tractable and that we had access to the gradient mapping $\theta \mapsto \nabla_\theta \log\ p(y|\theta)$. Recalling that Θ is a convex compact set, we could seek to iteratively solve (5) by using the projected gradient algorithm [24] which is given by $(\theta_n)_{n\in\mathbb{N}}$ with $\theta_0 \in \Theta$ and associated with the following recursion for any $n \in \mathbb{N}$

$$\theta_{n+1} = \Pi_\Theta \left[\theta_n + \delta \nabla_\theta \log\ p(y|\theta_n)\right] , \tag{8}$$

where Π_Θ is the projection onto Θ and $\delta > 0$ is a step-size. As mentioned previously, because in imaging problems $d_y \gg d_\Theta$, the marginal likelihood $\theta \mapsto p(y|\theta)$ typically exhibits a single maximiser $\theta_\star$ and is strongly log-concave w.r.t. θ in the neighbourhood of $\theta_\star$. Therefore we expect that (8) would quickly converge.

Since $\theta \mapsto \nabla_\theta \log\ p(y|\theta)$ is not tractable, we cannot directly use (8) to compute $\theta_\star$. However, we can replace $\theta \mapsto \nabla_\theta \log\ p(y|\theta)$ with a noisy estimate an consider a stochastic variant of the projected gradient algorithm. In particular, under mild assumptions using Fisher's identity see Proposition 3, (1) and the fact that for any $x \in \mathbb{R}^d$, $y \in \mathbb{R}^{d_y}$ and $\theta \in \Theta$, $p(x, y|\theta) = p(y|x)p(x|\theta)$, we have for any $\theta \in \Theta$

$$\nabla_\theta \log p(y|\theta) = \int_{\mathbb{R}^d} p(\tilde{x}|y,\theta) \nabla_\theta \log p(\tilde{x}, y|\theta) \mathrm{d}\tilde{x} = -\int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|y,\theta) \mathrm{d}\tilde{x} - \nabla_\theta \log(\mathrm{Z}(\theta)) . \tag{9}$$

Hence, we can use Monte Carlo Markov chain methods to approximate $\theta \mapsto \nabla_\theta \log p(y|\theta)$ for any $\theta \in \Theta$. We now consider a stochastic approximation proximal gradient algorithm (SAPG), see [37], where the expectation $\int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|y,\theta) \mathrm{d}\tilde{x}$ is replaced by a Monte Carlo estimator leading to the following gradient estimate for any $\theta \in \Theta$

$$\Delta_{m,\theta} = \frac{1}{m} \sum_{k=1}^{m} \nabla_\theta \log p(X_k, y|\theta) = -\frac{1}{m} \sum_{k=1}^{m} g(X_k) - \nabla_\theta \log \mathrm{Z}(\theta) ,$$

where $(X_k)_{k\in\{0,\ldots,m\}}$ is a sample of size $m \in \mathbb{N}^*$ generated by using a Markov Chain targeting $p(x|y,\theta) = p(x,y|\theta)/p(y|\theta)$, or a regularised approximation of this density. Therefore we can build a new sequence $(\theta_n)_{n\in\mathbb{N}}$ with θ_0 and associated with the following recursion for any $n \in \mathbb{N}$

$$\theta_{n+1} = \Pi_\Theta[\theta_n + \delta_{n+1} \Delta_{m_n,\theta_n}] , \qquad \Delta_{m_n,\theta_n} = -\frac{1}{m_n} \sum_{k=1}^{m_n} g(X_k^n) - \nabla_\theta \log \mathrm{Z}(\theta_n) , \tag{10}$$

where $(m_n)_{n\in\mathbb{N}}$ is a sequence of non-decreasing sample sizes and $(\delta_n)_{n\in\mathbb{N}}$ is a sequence of non-increasing step-sizes. Under some technical assumptions on the Markov kernels (see [6] for details), the errors in the gradient estimates asymptotically average out and the algorithm converges to a maximiser of $\theta \mapsto p(y|\theta)$. More precisely, given $N \in \mathbb{N}$ and a sequence $(\theta_n)_{n=0}^{N-1}$ generated using (10), an approximate solution of (5) can be obtained by calculating, for example, the average

$$\bar{\theta}_N = \frac{1}{N}\sum_{n=0}^{N-1} \theta_N, \tag{11}$$

which converges asymptotically to a solution of (5) as $N \to \infty$.

Applying this strategy to imaging problems is highly non-trivial because it requires generating very high-dimensional Markov chains $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ in a way that is computationally efficient and that satisfies a number of complex technical conditions on the associated Markov kernels (see Theorem 1 in Section 4). In this work, we address this major difficulty by constructing a SAPG scheme with state-of-the-art unadjusted proximal Markov kernel $\{\mathrm{R}_{\gamma,\theta} : \gamma \in (0,\bar{\gamma}],\ \theta \in \Theta\}$ that satisfy the required theoretical conditions, and where $\gamma \in (0,\bar{\gamma}]$ is an algorithm parameter that controls a trade-off between accuracy and computational efficiency, see Section 3.3. By doing so, we deliver a stochastic optimisation methodology that is highly computationally efficient and has strong theoretical guarantees.

New hyper param !

Lastly, observe that in order to use (10) it is necessary to evaluate $\theta \mapsto \nabla_\theta \log \mathrm{Z}(\theta)$. We propose three different strategies to address this calculation depending on whether g is a homogeneous function or not.

3.2.1 Scalar-valued θ with α positively homogeneous regulariser

For scalar-valued θ, *i.e.* $d_\Theta = 1$, (9) is given for any $\theta \in \Theta$ by

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(y|\theta) = -\int_{\mathbb{R}^d} g(x)p(\tilde{x}|y,\theta)\,\mathrm{d}\tilde{x} - \frac{\mathrm{d}}{\mathrm{d}\theta}\log \mathrm{Z}(\theta)\,. \tag{12}$$

Assume that there exists $\alpha \in \mathbb{R}\backslash\{0\}$ such that g is a α positively homogeneous function, *i.e.* for any $x \in \mathbb{R}^d$ and $t > 0$, $g(tx) = t^\alpha g(x)$, and recalling that $\Theta \subset (0,+\infty)$ we have for any $\theta \in \Theta$

$$\mathrm{Z}(\theta) = \int_{\mathbb{R}^d} \mathrm{e}^{-\theta g(\tilde{x})}\mathrm{d}\tilde{x} = \int_{\mathbb{R}^d} \mathrm{e}^{-g(\theta^{1/\alpha}\tilde{x})}\mathrm{d}\tilde{x} = \theta^{-d/\alpha}\int_{\mathbb{R}^d} \mathrm{e}^{-g(\tilde{x})}\mathrm{d}\tilde{x}\,,$$

and therefore

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log \mathrm{Z}(\theta) = -d/(\alpha\theta)\,.$$

Hence, (12) becomes for any $\theta \in \Theta$

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(y|\theta) = d/(\alpha\theta) - \int_{\mathbb{R}^d} g(x)p(\tilde{x}|y,\theta)\mathrm{d}\tilde{x}\,,$$

which leads to Algorithm 1 below. We want to point out that many commonly used regularisers are positively homogeneous. For example, all norms such as ℓ_1, ℓ_2, total variation (TV), nuclear or compositions of norms with linear operators (e.g., analysis terms of the form $\|\Psi x\|_1$, where $\Psi \in \mathbb{R}^{d_1}\times\mathbb{R}^d$ with $d_1 \in \mathbb{N}$) are 1 positively homogeneous. Moreover, powers of norms with exponent $q > 0$ are q positively homogeneous, and all linear combinations of positively homogeneous functions with the same homogeneity constant α, are also α positively homogeneous.

Algorithm 1 SAPG algorithm - Scalar θ and α positively homogeneous regulariser g

1: Input: initial $\{\theta_0, X_0\}$, $(\delta_n, \gamma_n)_{n\in\mathbb{N}}$, number of iterations N.
2: **for** $n = 0$ to $N-1$ **do**
3: Sample $X_{n+1} \sim \mathrm{R}_{\gamma_n,\theta_n}(X_n,\cdot)$,
4: Set $\theta_{n+1} = \Pi_\Theta\left[\theta_n + \delta_{n+1}\left(d/(\alpha\theta_n) - g(X_{n+1})\right)\right]$.
5: **end for**
6: Output: $\bar{\theta}_N = N^{-1}\sum_{n=0}^{N-1}\theta_n$.

3.2.2 Separably homogeneous regulariser

For the special case of separably homogeneous regularisers, Algorithm 1 can be adapted for multivariate θ. This is because in this class of regulariser, each component of θ affects independent subsets of the components of x. More precisely, assume that g is separably homogeneous in the following sense: there exist $(\tilde{g}_i)_{i\in\{1,\dots,d_\Theta\}}$, $(\mathsf{A}_i)_{i\in\{1,\dots,d_\Theta\}}$ pairwise disjoint subsets of $\{1,\dots,d\}$ and $(\alpha_i)_{i\in\{1,\dots,d_\Theta\}}$ such that for any $i \in \{1,\dots,d_\Theta\}$, $\tilde{g}_i : \mathbb{R}^{d_i} \to \mathbb{R}$ is α_i-positively homogeneous with $\alpha_i > 0$ and for any $x \in \mathbb{R}^d$, $g(x) = (\tilde{g}_i(x_{\mathsf{A}_i}))_{i\in\{1,\dots,d_\Theta\}}$ where for any $\mathsf{A} = \{i_1,\dots,i_\ell\} \subset \{1,\dots,d\}$, $x_{\mathsf{A}} = (x_{i_1},\dots,x_{i_\ell})$. In this case we have for any $\theta \in \Theta$

$$\begin{aligned}\mathrm{Z}(\theta) &= \int_{\mathbb{R}^d} \exp[-\theta^\top g(\tilde{x})]\mathrm{d}\tilde{x} = \int_{\mathbb{R}^d} \exp\left[-\sum_{i=1}^{d_\Theta} \theta^i \tilde{g}_i(\tilde{x}_{\mathsf{A}_i})\right] \mathrm{d}\tilde{x} \\ &= \prod_{i=1}^{d_\Theta} \int_{\mathbb{R}^{|\mathsf{A}_i|}} \exp[-\theta^i \tilde{g}_i(\tilde{x}_{\mathsf{A}_i})]\mathrm{d}\tilde{x} \,.\end{aligned}$$

Therefore, for any $i \in \{1,\dots,d_\Theta\}$ and $\theta \in \Theta$ we get that

$$[\partial \log \mathrm{Z}/\partial\theta^i](\theta) = -\left|\mathsf{A}_i\right|/(\alpha_i\theta^i).$$

Using this property we obtain Algorithm 2 below.

Algorithm 2 SAPG algorithm - Multivariate θ and separably homogeneous regulariser

Input: initial $\{\theta_0, X_0\}$, $(\delta_n, \gamma_n)_{n\in\mathbb{N}}$, number of iterations N.
for $n = 0$ to $N-1$ **do**
  Sample $X_{n+1} \sim \mathrm{R}_{\gamma_n,\theta_n}(X_n,\cdot)$,
  Set $\theta_{n+1} = \Pi_\Theta\left[\theta_n + \delta_{n+1}\left(\left|\mathsf{A}_i\right|/(\alpha_i\theta_n^i) - g(X_{n+1})\right)\right]$.
end for
Output: $\bar{\theta}_N = N^{-1}\sum_{n=0}^{N-1}\theta_n$.

For example, many works in the imaging literature adopt a so-called synthesis formulation where x represents the unknown image on some orthonormal wavelet basis $\Psi \in \mathbb{R}^{d\times d}$ with $J \in \mathbb{N}$ levels[1], and consider level-adapted ℓ_1 regularisations of the form

$$\theta^\top g(x) = \sum_{j=1}^{J} \theta_j \|x_{\mathsf{A}_j}\|_1$$

where x_{A_j} are the elements of x associated with the Jth level and $\theta \in \mathbb{R}^J$. Here, g is a separably homogeneous functional as it can be expressed as $g = (\tilde{g}_1,\dots,\tilde{g}_J)$ where, for any $j \in \{1,\dots,J\}$, $\tilde{g}_j$ is 1-positively homogeneous and $d_j = |\mathsf{A}_j|$. Notice that the domain in which x is represented is not relevant here; Algorithm 2 can be directly applied to any model where g is homogenous separable via a change of basis because the same expression for $\mathrm{Z}(\theta)$ holds.

3.2.3 General case: inhomogeneous regulariser

When g is neither homogeneous nor separably homogeneous, we address the evaluation of $\theta \mapsto \nabla_\theta \log \mathrm{Z}(\theta)$ numerically by stochastic simulation. More precisely, using that y is conditionally independent of θ given x, and using identity (2), we express $\theta \mapsto \nabla_\theta \log\, p(y|\theta)$ as the difference between two expectations, *i.e.* for any $\theta \in \Theta$

$$\nabla_\theta \log\, p(y|\theta) = \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta)\,\mathrm{d}\tilde{x} - \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|y,\theta)\,\mathrm{d}\tilde{x}\,.$$

We then use two families of Markov kernels $\{\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta} : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$ that respectively target the posterior $p(x|y,\theta)$ and the prior $p(x|\theta)$ within the SAPG Algorithm 3 below.

[1] In synthesis formulations $x \in \mathbb{R}^d$ represents the unknown image on some basis $\Psi \in \mathbb{R}^{d\times d}$; the solution in the pixel domain is given by $\Psi^\top x$.

Algorithm 3 SAPG algorithm - General form

Input: initial $\{\theta_0, X_0, \bar{X}_0\}$, $(\delta_n, \gamma_n, \gamma'_n)_{n\in\mathbb{N}}$, number of iterations N.
for $n = 0$ to $N-1$ **do**
  Sample $X_{n+1} \sim \mathrm{R}_{\gamma_n,\theta_n}(X_n, \cdot)$,
  Sample $\bar{X}_{n+1} \sim \bar{\mathrm{R}}_{\gamma'_n,\theta_n}(\bar{X}_n, \cdot)$,
  Set $\theta_{n+1} = \Pi_\Theta\left[\theta_n + \delta_{n+1}\left(g(\bar{X}_{n+1}) - g(X_{n+1})\right)\right]$.
end for
Output: $\bar{\theta}_N = N^{-1}\sum_{n=0}^{N-1}\theta_n$.

3.3 MCMC Kernels

Given the high dimensionality involved, it is fundamental to carefully choose the families of Markov kernels $\{\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ driving the SAPG. Here we use the MYULA Markov kernel recently proposed in [34], which is a state-of-the-art proximal Markov chain Monte Carlo (MCMC) method specifically designed for high-dimensional inverse problems that are convex but not smooth. This particular MCMC method is derived from the discretisation of an over-damped Langevin diffusion, $(\bar{X}_t)_{t\geqslant 0}$, satisfying the following stochastic differential equation

$$\mathrm{d}\mathbf{X}_t = -\nabla_x F(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t\,, \tag{13}$$

where $F : \mathbb{R}^d \mapsto \mathbb{R}$ is a continuously differentiable potential and $(\mathbf{B}_t)_{t\geqslant 0}$ is a standard d-dimensional Brownian motion. Under mild assumptions, this equation has a unique strong solution [43, Chapter 4, Theorem 2.3]. Accordingly, the law of $(X_t)_{t\geqslant 0}$ converges as $t \to \infty$ to the diffusion's unique invariant distribution, with probability density given by $\pi(x) \propto \mathrm{e}^{-F(x)}$ for all $x \in \mathbb{R}^d$ [75, Theorem 2.2]. Hence, to use (13) as a Monte Carlo method to sample from the posterior $p(x|y,\theta)$, we set $F(x) = -\nabla_x \log p(x|y,\theta)$ and thus specify the desired target density. Similarly, to sample from the prior we set $F(x) = -\nabla_x \log p(x|\theta)$.

However, sampling directly from (13) is usually not computationally feasible. Instead, we usually to resort to a discrete-time Euler-Maruyama approximation of (13) that leads to the following Markov chain $(X_k)_{k\in\mathbb{N}}$ with $X_0 \in \mathbb{R}^d$, given for any $k \in \mathbb{N}$ by

$$\mathrm{ULA} : X_{k+1} = X_k - \gamma\nabla_x F(X_k) + \sqrt{2\gamma}Z_{k+1},$$

where $\gamma > 0$ is a discretisation step-size and $(Z_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d d-dimensional zero-mean Gaussian random variables with an identity covariance matrix. This Markov chain is commonly known as the Unadjusted Langevin Algorithm (ULA) [75]. Under some additional assumptions on F, namely Lipschitz continuity of $\nabla_x F$, the ULA chain inherits the convergence properties of (13) and converges to a stationary distribution that is close to the target π, with γ controlling a trade-off between accuracy and convergence speed [34].

In this form, the ULA algorithm is limited to distributions where F is a Lipschitz continuously differentiable function. In the problems of interest this is very often not the case; when we sample from the posterior distribution $p(x|y,\theta)$ then for any $x \in \mathbb{R}^d$, $F(x) = f_y(x) + \theta^\top g(x)$ and when we sample from the prior distribution $x \mapsto p(x|\theta)$, for any $x \in \mathbb{R}^d$, $F(x) = \theta^\top g(x)$. In both cases, if g is not smooth then ULA is no longer applicable. The MYULA kernel was designed precisely to overcome this limitation.

Suppose that the target potential admits a decomposition $F = U + V$ where U is Lipschitz differentiable and V is not. In MYULA, the differentiable part is handled via the gradient $\nabla_x U$ in a manner to ULA, whereas the non-differentiable part is replaced by a smooth approximation $V^\lambda(x)$ given by the Moreau-Yosida envelope of V, see [11, Definition 12.20], defined for any $x \in \mathbb{R}^d$ and $\lambda > 0$ by

$$V^\lambda(x) = \min_{\tilde{x}\in\mathbb{R}^d}\left\{V(\tilde{x}) + (1/2\lambda)\left\|x - \tilde{x}\right\|_2^2\right\}\,. \tag{14}$$

For any $\lambda > 0$, the Moreau-Yosida envelope V^λ is continuously differentiable with gradient given for any $x \in \mathbb{R}^d$ by

$$\nabla V^\lambda(x) = (x - \mathrm{prox}_V^\lambda(x))/\lambda\,, \tag{15}$$

(see, e.g., [11, Proposition 16.44]). Using this approximation we obtain the MYULA kernel associated with $(X_k)_{k\in\mathbb{N}}$ given by $X_0 \in \mathbb{R}^d$ and the following recursion for any $k \in \mathbb{N}$

$$\mathrm{MYULA} : X_{k+1} = X_k - \gamma\nabla_x U(X_k) - \gamma\nabla_x V^\lambda(X_k) + \sqrt{2\gamma}Z_{k+1}\,.$$

Returning to the problem of interest, if we now choose to do the splitting such that $U = f_y$ and $V = \theta^\top g$, we can define the MYULA families of Markov kernels $\{\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ that we use in Algorithm 1, Algorithm 2 and Algorithm 3. For any $\theta \in \Theta$ and $\gamma > 0$, $\mathrm{R}_{\gamma,\theta}$ associated with $(X_k)_{k\in\mathbb{N}}$ given by $X_0 \in \mathbb{R}^d$ and the following recursion for any $k \in \mathbb{N}$

$$X_{k+1} = X_k - \gamma \nabla_x f_y(X_k) - \gamma \left\{ X_k - \mathrm{prox}^{\lambda}_{\theta^\top g}(X_k) \right\} / \lambda + \sqrt{2\gamma} Z_{k+1} \,, \tag{16}$$

where $\lambda > 0$. For any $\theta \in \Theta$ and $\gamma' > 0$, $\bar{\mathrm{R}}_{\gamma,\theta}$ associated with $(X_k)_{k\in\mathbb{N}}$ given by $X_0 \in \mathbb{R}^d$ and the following recursion for any $k \in \mathbb{N}$

$$\bar{X}_{k+1} = \bar{X}_k - \gamma' \left\{ \bar{X}_k - \mathrm{prox}^{\lambda'}_{\theta^\top g}(\bar{X}_k) \right\} / \lambda' + \sqrt{2\gamma} Z_{k+1} \,, \tag{17}$$

where we recall that $\lambda, \lambda' > 0$ are the smoothing parameters associated with $\theta^\top g^\lambda$, $\gamma, \gamma' > 0$ are the discretisation steps and $(Z_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d d-dimensional zero-mean Gaussian random variables with an identity covariance matrix.

We want to point out that this is not the only possible way of splitting $f_y + \theta^\top g$ into U and V. If some of the functions g_i are differentiable, it might be convenient to group those with f_y under U and leave only the non-differentiable terms in V. Moreover, doing the particular splitting we show in (16) requires computing the proximal operator of the global function $\theta^\top g$. In some cases it might be easier to use the proximal operators of each individual g_i independently. In this case, it is possible to replace each g_i with its smoothed version g_i^λ instead of doing it globally. Which choice is better will mostly depend on which tools are available to the practitioner. Note that most convex optimisation algorithms for MAP estimation (4) also use the operators ∇f_y and either $\mathrm{prox}^{\lambda}_{\theta^\top g}$ or $\mathrm{prox}^{\lambda}_{\theta^i g_i}$ [24, 41], making the implementation of Algorithm 3 straightforward for problems currently solved with such tools.

Finally, we note at this point that the MYULA kernels (16) and (17), do not target the posterior or prior distributions exactly but rather an approximation of these distributions. This is mainly due to two facts: 1) we are not able to use the exact Langevin diffusion (13), so we resort to a discrete approximation instead; and 2) we replace the non-differentiable terms with their Moreau-Yosida envelopes. As a result of these approximation errors, Algorithm 3 will exhibit some asymptotic estimation bias. This error is controlled by $\lambda, \lambda', \gamma, \gamma'$, and δ, and can be made arbitrarily small at the expense of additional computing time, see Theorem 2 in Section 4. The bias can also be completely removed by combining (16)-(17) with Metropolis-Hastings steps, as discussed in detail in [62]. However, preliminary experiments (not reported here) suggest that it is very difficult to calibrate high-dimensional Metropolis steps within a SAPG scheme to achieve the required acceptance rates, as the target densities change at each iteration, so we do not explore this any further. Also notice that incorporating the Metropolis-Hastings correction can significantly deteriorate non-asymptotic convergence properties in imaging problems, leading a significant increase in computing times [34].

3.4 Implementation guidelines

We now provide some practical guidelines for implementing Algorithm 3. These recommendations do not seek to define optimal values for specific models, but rather to provide general rules that are simple and robust. For more details we encourage the reader to see Appendix B where we have included additional practical recommendations for implementing this methodology and explain the effect of each algorithm parameter and the criteria for selecting their values.

In Algorithm 3 we recommend using the following values: $\lambda = 1/\mathrm{L}_y$, $\gamma_n = 0.98 \times (\mathrm{L}_y + 1/\lambda)^{-1}$ for any $n \in \mathbb{N}$, $\lambda' = \lambda$ and $\gamma'_n = 0.98 \times \lambda$ for any $n \in \mathbb{N}$. Regarding the choice of the sequence $(\delta_n)_{n\in\mathbb{N}}$, a standard choice is given for any $n \in \mathbb{N}^*$ by $\delta_n = c_0 n^{-p}/d$ for some $c_0 > 0$ and $p \in [0.6, 0.9]$. In our experiments we use $p = 0.8$. For c_0 we recommend, for the case where θ is scalar, to start with $c_0 = \theta_0^{-1}$ and then adjust if necessary. When θ is not scalar, it might be better to use different scales for every component of θ. For more details on this topic, see Appendix B.1. Lastly, we recommend implementing the algorithm including a *warm-up* initialisation during which the MCMC kernels should run with a fixed value of $\theta = \theta_0$ for a certain number of *warm-up* iterations T_0.

3.5 Connections to hierarchical Bayesian approaches

As we mentioned earlier, the Bayesian framework provides two main paradigms to select θ automatically: the empirical (already discussed in Section 3.1) and the hierarchical, which is currently the

predominant Bayesian approach in data science (see [65, 64] for examples in imaging sciences). We now discuss connections between the two paradigms and stress advantages and disadvantages.

In the hierarchical Bayesian paradigm, θ is modelled as an additional unknown quantity and it is assigned a prior distribution $p(\theta)$. This leads to an augmented posterior given for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$ by

$$p(x, \theta|y) = p(y|x, \theta)p(x|\theta)p(\theta)/p(y) \ .$$

There are two main ways of employing this augmented posterior. The first, and most popular, is to remove θ from the model by marginalisation, followed by inference on $x|y$ with the marginal posterior given for any $x \in \mathbb{R}^d$ by

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

The marginal posterior is then often used to perform minimum mean squared error (MMSE) estimation by computing

$$\hat{x}_{\mathrm{MMSE}} = \int_{\mathbb{R}^d} \tilde{x}\, p(\tilde{x}|y)\ \mathrm{d}\tilde{x}.$$

This can be achieved with a standard MCMC algorithm when $\mathrm{Z}(\theta)$ is tractable, e.g. Gibbs sampling, or with specialised algorithm that allows circumventing the evaluation of $\mathrm{Z}(\theta)$ at the expense of significant additional computational cost (see [65] for details). For some specific models it is also possible to compute an approximate marginal MMSE solution by using a deterministic variational Bayesian algorithm (e.g., see [7, 56]), but such algorithms have not yet been widely adopted because their implementation and performance remains very problem-specific. Alternatively, for the class of models considered in Section 3.2.1, one can also efficiently compute the marginal MAP estimator

$$\hat{x}_{\mathrm{MAP}} \in \underset{x \in \mathbb{R}^d}{\mathrm{argmin}}\ p(x|y) \ , \tag{18}$$

by using the majorisation-minimisation algorithm proposed in [64].

In order to understand the connection between this hierarchical Bayesian approach and the empirical Bayesian strategy used in this paper it is useful to express $p(x|y)$ as follows

$$p(x|y) = \int_{\Theta} p(x|y, \tilde{\theta})p(\tilde{\theta}|y)\mathrm{d}\tilde{\theta} \ ,$$

where we observe that $x \mapsto p(x|y)$ is effectively a weighted average of all the posteriors $x \mapsto p(x|y, \theta)$ parametrised by $\theta \in \mathbb{R}^{d_\Theta}$, with weights given by the marginal posterior $p(\theta|y)$, which represents the uncertainty in θ given the observed data y. If instead of $p(\theta|y)$ we perform the integration of $\theta \mapsto p(x|y, \theta)$ with respect to the Dirac distribution $\delta_{\theta_\star}$, we obtain the empirical Bayesian pseudo-posterior $x \mapsto p(x|y, \theta_\star)$ considered in this paper.

Note that in imaging problems the marginal posterior $p(\theta|y) \propto p(y|\theta)p(\theta)$ will be dominated by the marginal likelihood $p(y|\theta)$ because of the dimensionality of y. Therefore most of the mass of $p(\theta|y)$ will be close to $\theta_\star$. As a result, we expect that both the hierarchical and the empirical approaches will deliver broadly similar results. For models that are correctly specified both strategies should perform well, and hierarchical Bayes should moderately outperform empirical Bayes as it is decision-theoretically optimal [74].

However, most imaging models are over-simplistic and hence somewhat misspecified. Our experiments suggest that in this case the empirical Bayesian approach can outperform the hierarchical one. More precisely, we practically observe that the marginal posterior $p(\theta|y)$ typically has its maximum at a good value for θ, but struggles to concentrate and spreads its mass across a much wider range of values of θ. Consequently, $\theta \mapsto p(\theta|y)$ fails to sufficiently penalise poor models, which are given too much weight in $x \mapsto p(x|y)$ as a result. In this situation, the pseudo-posterior $x \mapsto p(x|y, \theta_\star)$ often delivers better inferences than the marginal posterior $x \mapsto p(x|y)$. In the context of inverse problems, this phenomenon is particularly clear in problems that are poorly conditioned and where the misspecification of the prior has a stronger effect on the inferences. This behaviour is observed in all the imaging problems reported in Section 5, and is particularly clear in the hyperspectral unmixing problem.

It is also worth mentioning at this point that there is another hierarchical Bayesian approach where x and θ are estimated jointly from y, without marginalisation [88, 64]. For example, one can perform joint MAP estimation

$$(\hat{x}_\star, \hat{\theta}_\star) = \underset{x \in \mathbb{R}^d,\ \theta \in \Theta}{\mathrm{argmax}}\ p(x, \theta|y) \ . \tag{19}$$

This strategy will generally produce mildly inferior results, except for the class of models considered in Section 3.2.1, where both hierarchical MAP estimation approaches (19) and (18) essentially produce the same results [64].

4 Analysis of the convergence properties

4.1 Notations and conventions

We denote by $\mathrm{B}(0,R)$ and $\overline{\mathrm{B}}(0,R)$ the open ball, respectively the closed ball, with radius R in $\mathbb{R}^d$. Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$, $\mathbb{F}(\mathbb{R}^d)$ the set of all Borel measurable functions on $\mathbb{R}^d$ and for $f \in \mathbb{F}(\mathbb{R}^d)$, $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{F}(\mathbb{R}^d)$ a μ-integrable function, denote by $\mu(f)$ the integral of f w.r.t. μ. For $f \in \mathbb{F}(\mathbb{R}^d)$, the V-norm of f is given by $\|f\|_V = \sup_{x\in\mathbb{R}^d} |f(x)|/V(x)$. Let ξ be a finite signed measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The V-total variation distance of ξ is defined as

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

If $V \equiv 1$, then $\|\cdot\|_V$ is the total variation norm on measures denoted by $\|\cdot\|_{\mathrm{TV}}$.

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

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

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

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

4.2 Main results

In this section we turn to the general optimisation problem of minimizing a function $f : \Theta \to \mathbb{R}$, with $\Theta \subset \mathbb{R}^{d_\Theta}$ with intractable gradients. This setting includes the problem of marginal maximum likelihood considered in 5). In this case, $f(\theta) = -\log p(y|\theta)$. We consider the following assumptions on f and Θ.

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

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

$$\|\nabla_\theta f(\theta_1) - \nabla_\theta f(\theta_2)\| \leqslant L_f \|\theta_1 - \theta_2\| .$$

A 3. *For any $\theta \in \Theta$, there exist $H_\theta, \bar{H}_\theta : \mathbb{R}^d \to \mathbb{R}^{d_\Theta}$ and two probability distributions $\pi_\theta, \bar{\pi}_\theta$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ satisfying for any $\theta \in \Theta$*

$$\nabla_\theta f(\theta) = \int_{\mathbb{R}^d} H_\theta(x) \mathrm{d}\pi_\theta(x) + \int_{\mathbb{R}^d} \bar{H}_\theta(x) \mathrm{d}\bar{\pi}_\theta(x) .$$

In addition, $(\theta, x) \mapsto H_\theta(x)$ and $(\theta, x) \mapsto \bar{H}_\theta(x)$ are measurable.

Note that the last assumption covers the cases discussed in Section 3.2

1) if the regulariser g is α positively homogeneous with $\alpha > 0$ and $d_\Theta = 1$, corresponding to Section 3.2.1 then for any $\theta \in \Theta$, $H_\theta = g$, $\bar{H}_\theta = -d/(\alpha\theta)$, π_θ is the probability measure with density w.r.t. the Lebesgue measure $x \mapsto p(x|y,\theta)$ and $\bar{\pi}_\theta$ is any probability measure ;

2) if the regulariser g is separably positively homogeneous as in Section 3.2.2, then for any $\theta \in \Theta$, $H_\theta = g$, $\bar{H}_\theta = (-|\mathsf{A}_i|/(\alpha_i \theta^i))_{i \in \{1,\ldots,d_\Theta\}}$, π_θ is the probability measure with density w.r.t. the Lebesgue measure $x \mapsto p(x|y,\theta)$ and $\bar{\pi}_\theta$ is any probability measure ;

3) if the regulariser g is inhomogeneous, corresponding to Section 3.2.3, then for any $\theta \in \Theta$, $\bar{H}_\theta = H_\theta = g$, π_θ and $\bar{\pi}_\theta$ are the probability measures with density w.r.t. the Lebesgue measure $x \mapsto p(x|y,\theta)$ and $x \mapsto p(x|\theta)$ respectively.

We now specify the probability distributions we are interested in. Note that the following assumption encompasses a large class of variational formulations in Bayesian imaging.

H1. *For any $\theta \in \Theta$, there exist $V_\theta, \bar{V}_\theta, U_\theta, \bar{U}_\theta : \mathbb{R}^d \to [0, +\infty)$ convex functions satisfying the following conditions.*

(a) For any $\theta \in \Theta$ and $x \in \mathbb{R}^d$,

$$\pi_\theta(x) \propto \exp\left[-V_\theta(x) - U_\theta(x)\right] , \qquad \bar{\pi}_\theta(x) \propto \exp\left[-\bar{V}_\theta(x) - \bar{U}_\theta(x)\right] ,$$

and

$$\min\left(\inf_{\theta\in\Theta} \int_{\mathbb{R}^d} \exp[-V_\theta(\tilde{x}) - U_\theta(\tilde{x})]\mathrm{d}\tilde{x}, \inf_{\theta\in\Theta} \int_{\mathbb{R}^d} \exp[-\bar{V}_\theta(\tilde{x}) - \bar{U}_\theta(\tilde{x})]\mathrm{d}\tilde{x}\right) > 0 . \tag{20}$$

(b) For any $\theta \in \Theta$, V_θ and $\bar{V}_\theta$ are continuously differentiable and there exists $\mathtt{L} \geqslant 0$ such that for any $\theta \in \Theta$ and $x, y \in \mathbb{R}^d$

$$\max\left(\|\nabla_x V_\theta(x) - \nabla_x V_\theta(y)\|, \|\nabla_x \bar{V}_\theta(x) - \nabla_x \bar{V}_\theta(y)\|\right) \leqslant \mathtt{L}\|x - y\| .$$

In addition, there exist $R_{V,1}, R_{V,2} \geqslant 0$ such that for any $\theta \in \Theta$, there exist $x^\star_\theta, \bar{x}^\star_\theta \in \mathbb{R}^d$ with $x^\star_\theta \in \arg\min_{\mathbb{R}^d} V_\theta$, $\bar{x}^\star_\theta \in \arg\min_{\mathbb{R}^d} \bar{V}_\theta$, $x^\star_\theta, \bar{x}^\star_\theta \in \overline{\mathrm{B}}(0, R_{V,1})$ and $V_\theta(x^\star_\theta), \bar{V}_\theta(\bar{x}^\star_\theta) \in \overline{\mathrm{B}}(0, R_{V,2})$.

(c) There exists $\mathtt{M} \geqslant 0$ such that for any $\theta \in \Theta$ and $x, y \in \mathbb{R}^d$

$$\max\left(\|U_\theta(x) - U_\theta(y)\|, \|\bar{U}_\theta(x) - \bar{U}_\theta(y)\|\right) \leqslant \mathtt{M}\|x - y\| .$$

In addition, there exist $R_{U,1}, R_{U,2} \geqslant 0$ such that for any $\theta \in \Theta$, there exist $x^\sharp_\theta, \bar{x}^\sharp_\theta \in \mathbb{R}^d$ with $x^\sharp_\theta, \bar{x}^\sharp_\theta \in \overline{\mathrm{B}}(0, R_{U,1})$ and $U_\theta(x^\sharp_\theta), \bar{U}_\theta(\bar{x}^\sharp_\theta) \in \overline{\mathrm{B}}(0, R_{U,2})$.

Note that (20) in **H1** (a) is satisfied if Θ is compact and the functions $\theta \mapsto \int_{\mathbb{R}^d} \exp[-V_\theta(\tilde{x}) - U_\theta(\tilde{x})]\mathrm{d}\tilde{x}$ and $\theta \mapsto \int_{\mathbb{R}^d} \exp[-\bar{V}_\theta(\tilde{x}) - \bar{U}_\theta(\tilde{x})]\mathrm{d}\tilde{x}$ are continuous. This latter condition can be then easily verified using the Lebesgue dominated convergence theorem and some assumptions on $\{V_\theta, \bar{V}_\theta, U_\theta, \bar{U}_\theta : \theta \in \Theta\}$. Note that if there exists $V : \mathbb{R}^d \to [0, +\infty)$ such that for any $\theta \in \Theta$, $V_\theta = V$ and there exists $x^\star \in \mathbb{R}^d$ with $x^\star \in \arg\min_{\mathbb{R}^d} V$ then one can choose $x^\star_\theta = x^\star$ for any $\theta \in \Theta$ in **H1** (b). In this case, $R_{V,2} = 0$. Similarly if for any $\theta \in \Theta$, $U_\theta(0) = 0$ then one can choose $x^\sharp_\theta = 0$ in **H 1** (c) and in this case $R_{U,1} = R_{U,2} = 0$.

As emphasized in Section 3.2 we use a stochastic approximation approach to minimize f and therefore need to consider Monte Carlo estimators for $\nabla_\theta f(\theta)$ and $\theta \in \Theta$ based on two MCMC schemes targeting π_θ and $\bar{\pi}_\theta$ respectively. Specifically, we consider two different but close MCMC methodologies to achieve such a task.

A first option, as proposed in Section 3.3, is to consider a Moreau-Yosida version of the Unadjusted Langevin Algorithm to sample from π_θ and $\bar{\pi}_\theta$. Let $\kappa > 1/2$ and $\{\mathrm{R}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ be the family of kernels defined for any $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$, $\theta \in \Theta$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{R}_{\gamma,\theta}(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(\left\| y - x + \gamma \nabla_x V_\theta(x) + \kappa^{-1}\left\{x - \mathrm{prox}^{\gamma\kappa}_{U_\theta}(x)\right\}\right\|^2 \Big/ (4\gamma) \right) \mathrm{d}y . \tag{21}$$

Consider also the family of Markov kernels $\{\bar{\mathrm{R}}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ such that for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, $\bar{\mathrm{R}}_{\gamma,\theta}$ is the Markov kernel defined by (21) but with $\bar{U}_\theta$ and $\bar{V}_\theta$ in place of U_θ and V_θ respectively. We highlight that κ is linked to λ in (16) in the following manner $\kappa = \lambda/\gamma$. In the previous section only MYULA kernels are considered, see (16) and (17). However, one can also use the following families of kernels, introduced in [31] and called Proximal Unadjusted Langevin Algorithm (PULA) in the

present document. Define the family $\{\mathrm{S}_{\gamma,\theta} : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$, for any $x \in \mathbb{R}^d$, $\gamma \in (0,\bar{\gamma}]$, $\theta \in \Theta$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{S}_{\gamma,\theta}(x,\mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(\left\| y - \mathrm{prox}^{\gamma\kappa}_{U_\theta}(x) + \gamma \nabla_x V_\theta(\mathrm{prox}^{\gamma\kappa}_{U_\theta}(x)) \right\|^2 \Big/ (4\gamma) \right) \mathrm{d}y \,. \tag{22}$$

Consider also the family of Markov kernels $\{\bar{\mathrm{S}}_{\gamma,\theta} : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$ such that for any $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \Theta$, $\bar{\mathrm{S}}_{\gamma,\theta}$ is the Markov kernel defined by (22) but with $\bar{U}_\theta$ and $\bar{V}_\theta$ in place of U_θ and V_θ respectively.

Starting from $(X_0^0, \bar{X}_0^0) \in \mathbb{R}^d \times \mathbb{R}^d$ and $\theta_0 \in \Theta$, we define on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the sequence $(\{(X_k^n, \bar{X}_k^n) : k \in \{0,\dots,m_n\}\}, \theta_n)_{n\in\mathbb{N}}$ by the following recursion for $n \in \mathbb{N}$ and $k \in \{0,\dots,m_n - 1\}$

$$\begin{aligned}
&(X_k^n)_{k\in\{0,\dots,m_n\}} \text{ is a MC with kernel } \mathrm{K}_{\gamma_n,\theta_n} \text{ and } X_0^n = X_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1} \,; \\
&(\bar{X}_k^n)_{k\in\{0,\dots,m_n\}} \text{ is a MC with kernel } \bar{\mathrm{K}}_{\gamma'_n,\theta_n} \text{ and } \bar{X}_0^n = \bar{X}_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1} \,; \\
&\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \left\{ H_{\theta_n}(X_k^n) + \bar{H}_{\theta_n}(\bar{X}_k^n) \right\} \right] \,,
\end{aligned} \tag{23}$$

where $\{(\mathrm{K}_{\gamma,\theta}, \bar{\mathrm{K}}_{\gamma,\theta}) : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\} = \{(\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta}) : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$ or $\{(\mathrm{S}_{\gamma,\theta}, \bar{\mathrm{S}}_{\gamma,\theta}) : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$, $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$, $\delta_n, \gamma_n, \gamma'_n > 0$ for any $n \in \mathbb{N}$, Π_Θ is the projection onto Θ and $\mathcal{F}_n$ is defined as follows for all $n \in \mathbb{N} \cup \{-1\}$

$$\mathcal{F}_n = \sigma\left(\theta_0, \{(X_k^\ell, \bar{X}_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}\right) \,, \qquad \mathcal{F}_{-1} = \sigma(\theta_0, X_0^0, \bar{X}_0^0) \,.$$

In order to analyse the sequence given by (23) we use the theoretical results derived in [26]. To this end, we impose that for any $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \Theta$, the kernels $\mathrm{K}_{\gamma,\theta}$ and $\bar{\mathrm{K}}_{\gamma,\theta}$ admit a invariant probability distribution, denoted by $\pi_{\gamma,\theta}$ and $\bar{\pi}_{\gamma,\theta}$ respectively which are approximations of $\bar{\pi}_\theta$ and $\bar{\pi}_\theta$ defined in A3 and geometrically converge towards them. Under one of the following assumptions we show that MYULA and PULA satisfy this kind of condition.

H2. *There exists $\mathtt{m} > 0$ such that for any $\theta \in \Theta$, V_θ and $\bar{V}_\theta$ are $\mathtt{m}$-convex.*

H3. *There exist $\eta > 0$ and $\mathtt{c} \geqslant 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\min(U_\theta(x), \bar{U}_\theta(x)) \geqslant \eta \|x\| - \mathtt{c}$.*

Note that if $\sup_{\theta\in\Theta}(\int_{\mathbb{R}^d} \exp[-U_\theta(\tilde{x})]\mathrm{d}\tilde{x}) < +\infty$, then H3 is automatically satisfied, as an immediate extension of [10, Lemma 2.2 (b)]. The following assumption ensures that the Markov kernels we consider are Lipschitz with respect to their parameters.

H4. *There exist $\mathtt{M}_\Theta \geqslant 0$ and $\mathtt{f}_\Theta \in \mathrm{C}(\mathbb{R}_+, \mathbb{R}_+)$ such that for any $\theta_1, \theta_2 \in \Theta$, $x \in \mathbb{R}^d$,*

$$\begin{aligned}
&\max\left(\|\nabla_x V_{\theta_1}(x) - \nabla_x V_{\theta_2}(x)\|, \|\nabla_x \bar{V}_{\theta_1}(x) - \nabla_x \bar{V}_{\theta_2}(x)\|\right) \leqslant \mathtt{M}_\Theta \|\theta_1 - \theta_2\| (1 + \|x\|) \,, \\
&\max\left(\|\nabla_x U^\kappa_{\theta_1}(x) - \nabla_x U^\kappa_{\theta_2}(x)\|, \|\nabla_x \bar{U}^\kappa_{\theta_1}(x) - \nabla_x \bar{U}^\kappa_{\theta_2}(x)\|\right) \leqslant \mathtt{f}_\Theta(\kappa) \|\theta_1 - \theta_2\| (1 + \|x\|) \,.
\end{aligned}$$

We now state our main results. In Theorem 1, we give sufficient conditions on the parameters of the algorithm under which the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s., and we give explicit convergence rates in Theorem 2. Consider for any $m \in \mathbb{N}^*$ and $\alpha > 0$, the two functions W_m and W_α given for any $x \in \mathbb{R}^d$ by

$$W_m(x) = 1 + \|x\|^{2m} \,, \qquad W_\alpha = \exp\left[\alpha\sqrt{1 + \|x\|^2}\right] \,.$$

Theorem 1. *Assume A1, A2, A3 and that f is convex. Let $\kappa > 1/2$. Assume H1 and one of the following condition:*

(a) H2 holds, $\bar{\gamma} < \min(2/(\mathtt{m} + \mathtt{L}), (2 - 1/\kappa)/\mathtt{L}, \mathtt{L}^{-1})$ and there exists $m \in \mathbb{N}^*$ and $C_m \geqslant 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leqslant C_m W_m^{1/4}(x)$ and $\|\bar{H}_\theta(x)\| \leqslant C_m W_m^{1/4}(x)$.

(b) H3 holds, $\bar{\gamma} < \min((2 - 1/\kappa)/\mathtt{L}, \eta/(2\mathtt{ML}), \mathtt{L}^{-1})$ and there exists $0 < \alpha < \eta/4$, $C_\alpha \geqslant 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leqslant C_\alpha W_\alpha^{1/4}(x)$ and $\|\bar{H}_\theta(x)\| \leqslant C_\alpha W_\alpha^{1/4}(x)$.

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 non-decreasing positive integers satisfying $\delta_0 < 1/L_f$ and $\gamma_0 < \bar{\gamma}$. Let $(\{(X_k^n, \bar{X}_k^n) : k \in \{0,\dots,m_n\}\}, \theta_n)_{n\in\mathbb{N}}$ be given by (23). In addition, assume that $\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty$, $\sum_{n=0}^{+\infty} \delta_{n+1}\gamma_n^{1/2} < +\infty$ and that one of the following condition holds:

(1) $\sum_{n=0}^{+\infty} \delta_{n+1}/(m_n\gamma_n) < +\infty$;

(2) $m_n = m_0 \in \mathbb{N}^*$ *for all* $n \in \mathbb{N}$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n| \, \delta_n^{-2} < +\infty$, **H4** *holds and we have* $\sum_{n=0}^{+\infty} \delta_{n+1}^2 \gamma_n^{-2} < +\infty$, $\sum_{n=0}^{+\infty} \delta_{n+1}\gamma_{n+1}^{-5/2}(\gamma_n - \gamma_{n+1})^{1/2} < +\infty$.

Then $(\theta_n)_{n\in\mathbb{N}}$ *converges a.s. to some* $\theta_\star \in \arg\min_\Theta f$. *Furthermore, a.s. there exists* $C \geqslant 0$ *such that for any* $n \in \mathbb{N}^*$

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

Proof. The proof is postponed to Appendix D. □

These results are similar to the ones identified in [80, Theorem 1, Theorem 5, Theorem 6] for the Stochastic Optimization with Unadjusted Langevin (SOUL) algorithm. Note that in SOUL the potential is assumed to be differentiable and the sampler is given by ULA, whereas in Theorem 1 the results are stated for PULA and MYULA samplers. As announced in [47], quantitative bounds can be derived in the non convex case. However, we leave it as a future study to extend our results to the case where f is non convex.

Assume that $\delta_n \sim n^{-a}$, $\gamma_n \sim n^{-b}$ and $m_n \sim n^{-c}$ with $a, b, c \geqslant 0$. We now distinguish two cases depending on if for all $n \in \mathbb{N}$, $m_n = m_0 \in \mathbb{N}^*$ (fixed batch size) or not (increasing batch size).

1) In the increasing batch size case, Theorem 1 ensures that $(\theta_n)_{n\in\mathbb{N}}$ converges if the following inequalities are satisfied

$$a + b/2 > 1 , \qquad a - b + c > 1 , \qquad a \leqslant 1 . \tag{24}$$

Note in particular that $c > 0$, *i.e.* the number of Markov chain iterates required to compute the estimator of the gradient increases at each step. However, for any $a \in [0, 1]$ there exist $b, c > 0$ such that (24) is satisfied. In the special setting where $a = 0$ then for any $\varepsilon_2 > \varepsilon_1 > 0$ such that $b = 2 + \varepsilon_1$ and $c = 3 + \varepsilon_2$ satisfy the results of (24) hold.

2) In the fixed batch size case, which implies that $c = 0$, Theorem 1 ensures that $(\theta_n)_{n\in\mathbb{N}}$ converges if the following inequalities are satisfied

$$a + b/2 > 1 , \qquad 2(a - b) > 1 , \qquad a + (b + 1)/2 - 5b/2 > 1 \qquad a \leqslant 1 ,$$

which can be rewritten as

$$b \in (2(a - 1), \min(a - 1/2, a/2 - 1/4)) , \qquad a \in [0, 1] .$$

The interval $(2(a - 1), \min(a - 1/2, a/2 - 1/4))$ is then not empty if and only if $a \in (9/10, 1]$.

Theorem 2. *Assume* **A 1**, **A 2**, **A 3** *and that* f *is convex. Let* $\kappa > 1/2$. *Assume* **H 1** *and that the condition (a) or (b) in Theorem 1 is satisfied. 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 non-decreasing positive integers satisfying* $\delta_0 < 1/L_f$ *and* $\gamma_0 < \bar{\gamma}$. *Let* $(\{(X_k^n, \bar{X}_k^n) : k \in \{0, \dots, m_n\}\}, \theta_n)_{n\in\mathbb{N}}$ *be given by* (23)

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

where

(a)

$$E_n = C_1 \left\{ 1 + \sum_{k=0}^{n-1} \delta_{k+1}\gamma_k^{1/2} + \sum_{k=0}^{n-1} \delta_{k+1}/(m_k\gamma_k) + \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 \right\} . \tag{25}$$

(b) or if $m_n = m_0$ *for all* $n \in \mathbb{N}$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n| \, \delta_n^{-2} < +\infty$ *and* **H4** *holds*

$$E_n = C_2 \left\{ 1 + \sum_{k=0}^{n-1} \delta_{k+1}\gamma_k^{1/2} + \sum_{k=0}^{n-1} \delta_{k+1}^2/\gamma_k^2 + \sum_{k=0}^{n-1} \delta_{k+1}\gamma_{k+1}^{-5/2}(\gamma_k - \gamma_{k+1})^{1/2} \right\} . \tag{26}$$

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

First, note that if the stepsize is fixed and recalling that $\kappa = \lambda/\gamma$ then the condition $\gamma < (2-1/\kappa)/\mathrm{L}$ can be rewritten as $\gamma < 2/(\mathrm{L} + \lambda^{-1})$. Assume that $(\delta_n)_{n\in\mathbb{N}}$ is non-increasing, $\lim_{n\to+\infty} \delta_n = 0$, $\lim_{n\to+\infty} m_n = +\infty$ and $\gamma_n = \gamma_0 > 0$ for all $n \in \mathbb{N}$. In addition, assume that $\sum_{n\in\mathbb{N}^*} \delta_n = +\infty$ then, by [70, Problem 80, Part I], it holds that

$$\begin{cases} \lim_{n\to+\infty} \left[\left(\sum_{k=1}^n \delta_k/m_k\right) / \left(\sum_{k=1}^n \delta_k\right) \right] = \lim_{n\to+\infty} 1/m_n = 0 \; ; \\ \lim_{n\to+\infty} \left[\left(\sum_{k=1}^n \delta_k^2\right) / \left(\sum_{k=1}^n \delta_k\right) \right] = \lim_{n\to+\infty} \delta_n = 0 \; . \end{cases} \tag{27}$$

Therefore, using (25) we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \Big/ \sum_{k=1}^n \delta_k \right\} - \min f \right] \leqslant C_1 \sqrt{\gamma_0} \; .$$

Similarly, if the stepsize is fixed and the number of Markov chain iterates is fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma_0$ and $m_n = m_0$ with $\gamma_0 > 0$ and $m_0 \in \mathbb{N}^*$, combining (26) and (27) we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \Big/ \sum_{k=1}^n \delta_k \right\} - \min f \right] \leqslant C_2 \sqrt{\gamma_0} \; .$$

5 Numerical experiments

In this section we validate the proposed methodology with a range of imaging inverse problems, which we have selected to illustrate a variety of observation models and regularisation functions. The first experiment, presented in Section 5.1, is carried out using synthetic images for which the exact generative statistical model is known, as this allows assessing the performance of the proposed method in a case where the regularisation parameter has a true value, and where there is no model misspecification. Following on from this, in Section 5.2 we demonstrate the method by estimating a scalar-valued regularisation parameter in a non-blind image deconvolution model with different kinds of prior distributions, such as total variation and ℓ_1-wavelet priors. This allows comparing our method to some state-of-the-art approaches that are limited to scalar-valued regularisation parameters. This is then followed by two challenging problems involving multivariate regularisation parameters. In particular, in Section 5.3 we apply our method to a sparse hyperspectral unmixing problem combining an ℓ_1 and a total variation regularisation, and where we report comparisons with the hierarchical Bayesian approach of [64]. Lastly, in Section 5.4 we apply our method to a total generalised variation denoising model that has two unknown regularisation parameters exhibiting strong dependencies, and which requires using Algorithm 3 with two parallel Markov chains.

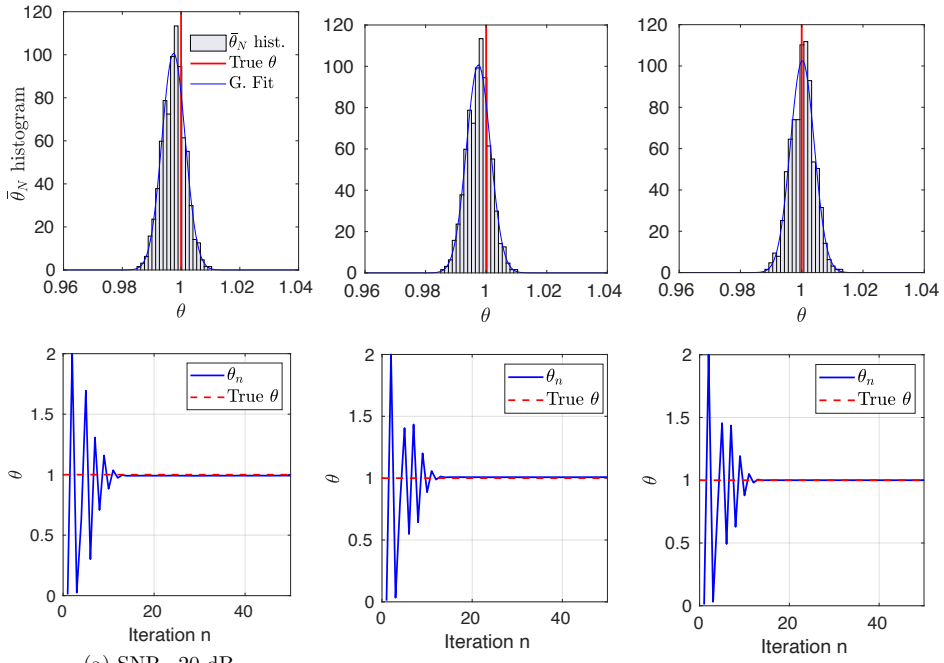
In all the experiments we first compute $\bar{\theta}_N$, see (11), and then calculate a MAP estimator using the empirical Bayesian posterior $x \mapsto p(x|y, \bar{\theta}_N)$ by convex optimisation (solver details are provided in each experiment). All experiments were conducted on an Intel i9-8950HK@2.90GHz workstation running Matlab R2018a.

5.1 Performance on synthetic images - denoising

We first demonstrate the performance of the algorithm on a very simple image denoising problem, where we work with synthetic test images to have access to the true value of the regularisation parameter. We consider a wavelet-based image denoising under a synthesis formulation where we assume that the coefficients x of the true image in an orthogonal 4-level Haar basis Ψ follow a Laplace distribution. That is the model (3) is given for any $x \in \mathbb{R}^d$ by $f_y(x) = \|y - \Psi x\|_2^2/(2\sigma^2)$ and $g(x) = \|x\|_1$. In our experiments, y has dimension $d_y = 256 \times 256$ pixels, and we set $\theta = 1$ to generate the synthetic test images. The variance of the added noise σ^2 is chosen for every case such that the signal-to-noise-ratio (SNR) is 20 dB, 30 dB, or 40 dB. In all cases we compute the empirical Bayes estimator $\bar{\theta}_N$ by implementing Algorithm 1 using the MYULA kernel (16).

To study the statistical behaviour of the method, we repeat each experiment 500 times by we generating 500 random observations y, each one coming from a different random x; then, for every

observation y, we estimate $\bar{\theta}_N$. Figure 2 shows the histograms obtained from the 500 estimated $\bar{\theta}_N$ values for each experiment (20 dB, 30 dB, and 40 dB). For completeness, we also present in Figure 2 one example of a generated sequence of iterates $(\theta_n)_{n\in\mathbb{N}}$ for each experiment. Observe that the estimation error is close to Gaussian and close to the true value of the regularisation parameter, as expected for a maximum likelihood estimator. The algorithm converges in approximately 15 iterations, possibly with some very small bias of the order of 0.1%.



(a) SNR=20 dB (b) SNR=30 dB (c) SNR=40 dB

Figure 2: Denoising with synthesis-ℓ_1 prior. Histograms of the estimated $\bar{\theta}_N$ for 500 repetitions and evolution of the iterates $(\theta_n)_{n\in\mathbb{N}}$. Results for SNR of 20 dB, 30 dB and 40 dB.

5.2 Non-blind natural image deconvolution

We now illustrate the proposed methodology with an application to image deblurring using two different kinds of prior distributions: the total variation (TV) prior and a wavelet-based synthesis-ℓ_1 prior. For comparison, we also report the results obtained with SUGAR [27] (only when using a TV prior), joint MAP estimation [64], discrepancy principle [36, 60], and by using the oracle value $\theta_\dagger$ that minimises the estimation mean squared error (MSE), *i.e.*

$$\theta_\dagger = \arg\min_{\theta\in\Theta} \left\{ \left\| x^0 - \arg\max_{x\in\mathbb{R}^d} p(x|y,\theta) \right\|_2 \right\},$$

where x^0 is the ground-truth. We want to highlight that carrying out such a comparison is not a trivial task because some algorithms are solver-dependent while some others are completely independent of the solver used to compute the MAP estimator. For this reason the comparison was done with extreme care, and we include a detailed explanation of how we compare the results in Appendix C.

In non-blind image deblurring, the aim is to recover an unknown image $x \in \mathbb{R}^d$ from a blurred and noisy observation $y = \mathrm{A}x + \mathrm{w}$, where $\mathrm{A} \in \mathbb{R}^d \times \mathbb{R}^d$ is a blur matrix, and w is a d-dimensional Gaussian random variable with zero mean and covariance matrix $\sigma\,\mathrm{Id}$ with $\sigma > 0$. In our experiments, x and y are of size $d = 512 \times 512$ pixels, A implements a known circulant uniform blur of size 9×9 pixels, and

σ^2 is chosen such that the blurred signal-to-noise-ratio (SNR) is 20 dB, 30 dB, or 40 dB. We perform all experiments on ten standard test images (`barbara, boat, bridge, flintstones, goldhill, lake, lena, man, mandrill and wheel`).

For each image, noise level, and θ selection method, we first obtain an estimate for θ and then use it to compute the MAP estimator $\hat{x}_{\mathrm{MAP}}$ (given by (4)). In the case of the joint MAP method [64], we carry out joint MAP estimation of θ and $\hat{x}_{\mathrm{MAP}}$. We compute the MAP estimator by using a highly efficient proximal convex optimisation algorithm, SALSA [1], which is an instance of Alternative Direction Method of Multipliers (ADMM). We then assess the resulting performance by computing the MSE between the MAP estimator and the ground truth.

5.2.1 Deconvolution with Total Variation prior

In this experiment we use model (3) where for any $x \in \mathbb{R}^d$ we have $f_y(x) = \|y - \mathrm{A}x\|_2^2 / 2\sigma^2$, $g(x) = \mathrm{TV}(x)$, and follow the previously explained procedure. Here $\mathrm{TV}(x)$ is the isotropic total variation pseudo-norm given by $\mathrm{TV}(x) = \sum_i \sqrt{(\Delta_i^h x)^2 + (\Delta_i^v x)^2}$ where Δ_i^v and Δ_i^h denote horizontal and vertical first-order local difference operators. To compute θ_N we use Algorithm 1. The prior associated with the total variation pseudonorm is not proper, therefore the effective dimension is $d - 1$. We evaluated the proximal operator of $\mathrm{TV}(x)$ using the primal-dual algorithm from [20] with 25 iterations.

The algorithm parameters are chosen following the recommendations provided in Appendix B.1; we consider 300 warm-up iterations and set $\theta_0 = 0.01$, $X_0 = y$, $\delta_n = 0.1 \times n^{-0.8}/d$ for any $n \in \mathbb{N}^*$, and we set $\lambda = \min\left(5L^{-1}, \lambda_{\max}\right)$ with $\lambda_{\max} = 2$ and $L = (0.99/\sigma)^2$. Since the algorithm converges rapidly, the first 25 values of $(\theta_n)_{n\in\mathbb{N}}$ are discarded before taking the average, that is, we consider for any $N_0 = 25$ and for any $N \in \mathbb{N}$ with $N \geqslant N_0$

$$\bar{\theta}_N = (N - N_0)^{-1} \sum_{i=N_0}^{N-1} \theta_i \,. \tag{28}$$

In addition, instead of setting a fixed number of iterations, we stop the algorithm when the relative change $|\bar{\theta}_{N+1} - \bar{\theta}_N|$ is smaller than 10^{-3}. It would be possible to use a tolerance of 10^{-5} and get a slight improvement of the MSE (< 0.02 dB), but this would lead to computing times that are five times longer. We use SALSA with the following parameters: `inneriters` $= 1$, `outeriters` $= 500$, `tol` $= 10^{-5}$ and `mu` $= \bar{\theta}_N/10$.

For illustration, Figure 3 shows the results obtained for two of the test images (`man` and `goldhill`) using the proposed method. The displayed images correspond to the 30 dB SNR setup. In Figure 4 we compare the MAP estimates obtained by using each of the considered methods. In this case we display a close-up on `man` and `goldhill` selecting a region that contains fine details and sharp edges. In Figure 5 and Figure 6 we provide further details for the same two images, showing a plot of the MSE obtained with each method and the evolution of the iterates $(\theta_n)_{n\in\mathbb{N}}$ for the empirical Bayesian method.

Observe in Figure 5 that the proposed empirical Bayesian algorithm yields close-to-optimal results, both for high and low SNR values. The method based on the discrepancy principle and the hierarchical Bayesian method overestimate the amount of regularisation required. Conversely, SUGAR underestimates θ (this can also be observed in the recovered image in Figure 4 (f), where the MAP estimate presents some ringing artefacts due to high-frequency noise amplification); this is in agreement with the results reported in [55].

Table 1 reports the average MSE values and average computing times obtained for each method. We can see that the proposed method performs close to the oracle performance, generally outperforming the other approaches from the state of the art with very competitive computing times. In particular, observe that the proposed method performs remarkably for all SNR values. At high SNR values (40 dB) discrepancy principle and joint MAP [64] perform similarly, whereas for low SNR values (20 dB) discrepancy principle outperforms joint MAP. Also, SUGAR performs well for low SNR, but fails to find good values of θ when the SNR is higher. This might be due to the fact that SUGAR minimises a surrogate of the MSE that works well for denoising but degrades in problems that are ill-posed or ill-conditioned. We emphasise at this point that the exact computing times of each algorithm depend on the specific stopping criteria and implementation details, so rather than claiming that one method is faster than the others, what we wish to illustrate is that the computing times are all within the same order of magnitude, with SUGAR being moderately slower for this particular experiment. As we mentioned before, if we had selected a tolerance of 10^{-5} to stop our algorithm, the computing

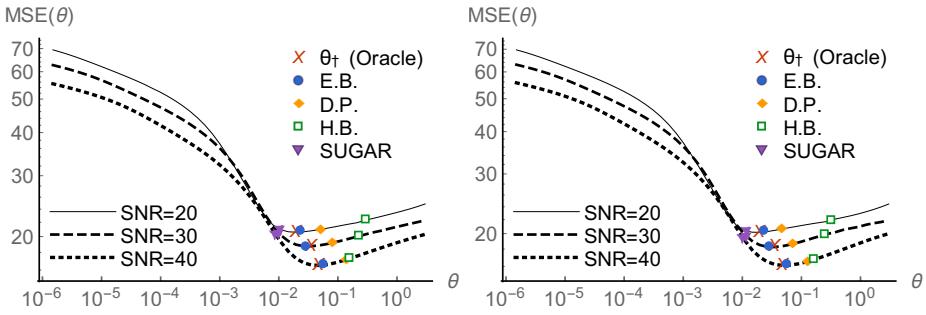
(a) Degraded (b) EB

Figure 3: Deblurring with TV prior for `man` and `goldhill` test images: (a) blurred and noisy (SNR=30 dB) observation y, (b) MAP estimator obtained using $\bar{\theta}_N$ computed with empirical Bayes.

(a) Original (b) Degraded (c) EB (d) HB (e) DP (f) SUGAR

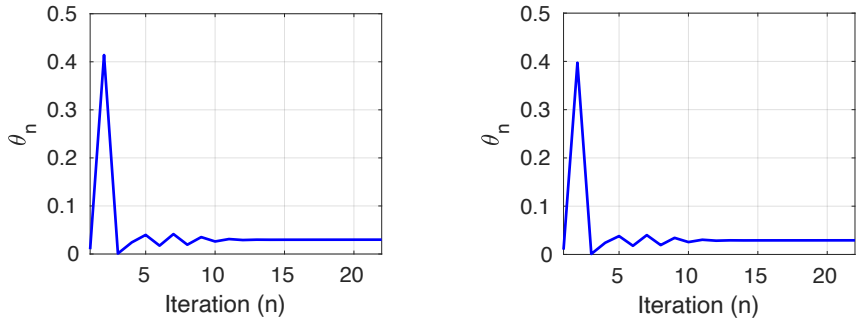
Figure 4: Deblurring with TV prior. Close-up on `man` and `goldhill` test images: (a) True image x, (b) blurred and noisy (SNR=30 dB) observation y, (c)-(f) MAP estimators obtained through empirical Bayes, hierarchical Bayes, discrepancy principle and SUGAR methods respectively.

times would have increased with almost negligible changes in the MSE. Also note that we compute the optimal θ for the discrepancy principle method by continuation, but one could also use a different



(a) `man` (b) `goldhill`

Figure 5: Deblurring with TV prior. Mean squared error (MSE) obtained for (a) `man` and (b) `goldhill` for different values of θ. We compare the values obtained with empirical Bayes, discrepancy principle, hierarchical Bayes, SUGAR, and the optimal value $\theta_\dagger$ that minimizes the MSE.



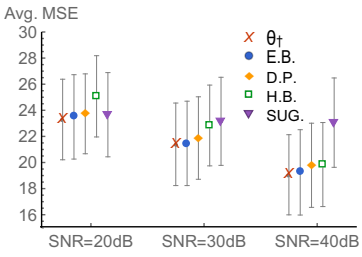
(a) `man` (b) `goldhill`

Figure 6: Deblurring with TV prior. Evolution of the sequence of iterates $(\theta_n)_{n\in\mathbb{N}}$ for the proposed method for `man` and `goldhill` test images (SNR=30 dB).

proximal splitting strategy (see [24] for instance).

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	MSE	Time	MSE	Time	MSE	Time
$\theta_\dagger$	23.29		21.39		19.06	
E.B.	23.50	0.84	21.46	0.85	19.24	0.85
D.P.	23.73	0.70	21.87	1.52	19.78	3.92
H.B.	25.07	0.58	22.84	1.27	19.84	3.27
SUGAR	23.66	3.64	23.16	5.00	23.05	5.63

(a)



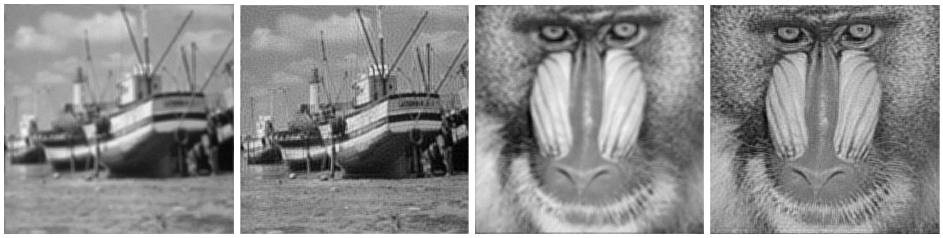
(b)

Table 1: Deblurring with TV prior. (a) Table with average mean squared error obtained for ten images with different algorithms. Average execution times expressed in minutes. In (b) we summarize the content of the table and show the standard deviation with error bars.

5.2.2 Wavelet deconvolution with synthesis prior

We now consider image deblurring under a wavelet synthesis formulation, where we assume that $x \in \mathbb{R}^d$ represents the unknown image in an redundant 4-level Haar wavelet representation Ψ, with dimension $d = 10 \times d_y = 10 \times 512 \times 512$ coefficients. We assume a Laplace prior on the elements of x with unknown parameter θ. Accordingly, the posterior is of the form (3) with $f_y(x) = \|y - \mathrm{A}\Psi x\|_2^2/(2\sigma^2)$, $g(x) = \|x\|_1$. To obtain solutions we map x to pixel domain by computing $\Psi^\top x$.

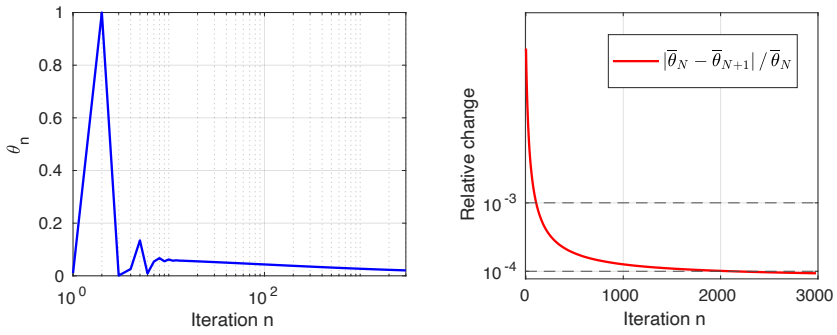
To compute $\bar{\theta}_N$ we use Algorithm 1. The algorithm parameters are chosen following the recommendations provided in Appendix B.1, we do not consider any warm-up iterations, and set $\theta_0 = 0.01$, $X_0 = y$, for any $n \in \mathbb{N}^*$, $\delta_n = 10 \times n^{-0.8}/d$, $\lambda = \min\left(5L^{-1}, \lambda_{\max}\right)$ with $\lambda_{\max} = 2$ and $L = (0.98/\sigma)^2$. We use the same stopping criteria as in the previous experiment and we consider two different tolerance levels: i) we stop the algorithm when the relative change $|\theta_{N+1} - \theta_N|$ is smaller than 10^{-4}, and ii) when the relative change is smaller than 10^{-3}. As in the previous experiment, we skip the first 20 iterations before computing the average $\bar{\theta}_N$. To compute the MAP estimate we use SALSA with the following parameters: `inneriters` $= 1$, `outeriters` $= 1000$, `tol` $= 10^{-5}$ and `mu` $= \bar{\theta}_N$.



(a) Degraded (b) EB (c) Degraded d) EB

Figure 7: Wavelet deconvolution with synthesis-ℓ_1 prior for `boat` and `mandrill` test images: (a),(c) blurred and noisy (SNR=20 dB) observation y, (b),(d) MAP estimator obtained with empirical Bayes.

In Figure 7 we show the results obtained for two of the test images (`boat` and `mandrill`) using the proposed method. The displayed images correspond to the 20 dB SNR setup. In Figure 8 we provide further details for the `boat` image, showing the evolution of the iterates $(\theta_n)_{n\in\mathbb{N}}$ and the relative differences on its running average value $(\bar{\theta}_N)_{N\in\mathbb{N}}$ throughout iterations.

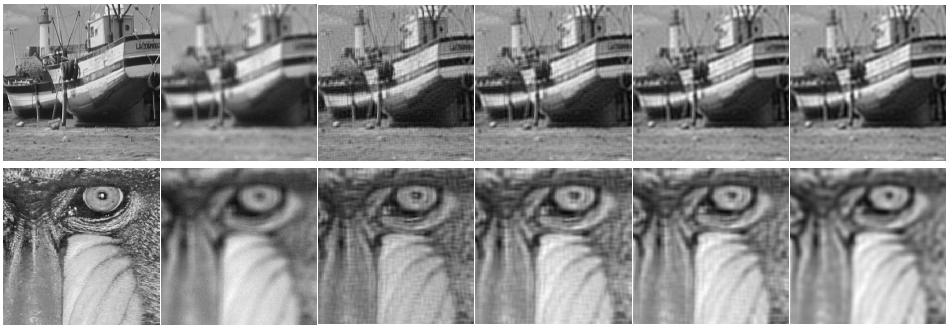


(a) (b)

Figure 8: Wavelet deconvolution with synthesis-ℓ_1 prior for `boat` image (SNR=20 dB). Evolution of (a) the iterates $(\theta_n)_{n\in\mathbb{N}}$ in log-scale and (b) the relative change in $(\bar{\theta}_N)_{N\in\mathbb{N}}$ for the proposed method.

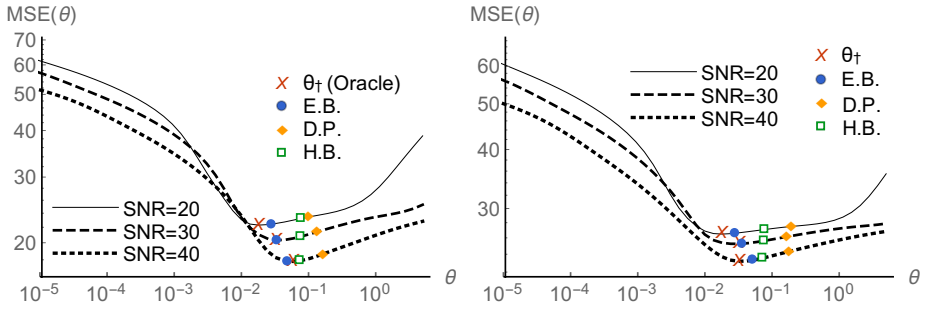
In Figure 9 we compare the results obtained by using each of the considered methods, showing a close-up on an image region that contains fine details and sharp edges. Figure 10 shows a plot of the MSE obtained with each method for the same two test images.

Table 2 shows the average MSE values and average computing times obtained for each method.



(a) Original (b) Degraded (c) EB $_{\text{tol } 10^{-4}}$ (d) EB $_{\text{tol } 10^{-3}}$ (e) HB (f) DP

Figure 9: Wavelet deconvolution with synthesis-ℓ_1 prior. Close-up on `boat` and `mandrill` images: (a) True image, (b) blurred and noisy (SNR=20 dB) observation y, (c)-(f) MAP estimators obtained with Empirical Bayes (tol. 10^{-4} and 10^{-3}), hierarchical Bayes and discrepancy principle respectively.



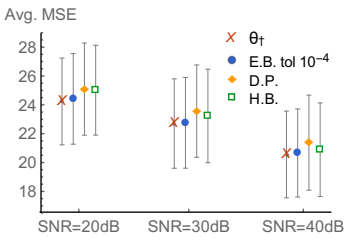
(a) `boat` (b) `mandrill`

Figure 10: Wavelet deconvolution with synthesis-ℓ_1 prior - Mean squared error (MSE) obtained for (a) `boat` and (b) `mandrill` for different values of θ. We compare the values obtained with empirical Bayes with tolerance 10^{-4}, discrepancy principle, hierarchical Bayes, and the optimal value $\theta_\dagger$.

We observe once again that the empirical Bayesian method achieves the best results for all SNR values and is very close to the oracle performance. Reducing the tolerance leads to a small improvement in MSE, at the expense of a higher computing time. The discrepancy principle consistently overestimates the parameter leading to over-smoothed solutions.

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	MSE	Time	MSE	Time	MSE	Time
$\theta_\dagger$	24.23		22.70		20.56	
$_{tol\ 10^{-4}}$ E.B.	24.40	4.48	22.80	3.59	20.70	2.44
$_{tol\ 10^{-3}}$ E.B.	24.70	0.36	22.90	0.25	20.80	0.09
D.P.	25.09	13.93	23.57	28.64	21.38	61.03
H.B.	25.01	11.61	23.23	23.87	20.89	50.86

(a)



(b)

Table 2: Wavelet deconvolution with synthesis-ℓ_1 prior. (a) Table with average mean squared error obtained for ten images with different algorithms. Average execution times expressed in minutes. (b) summarises the content of the table and shows the standard deviation with error bars.

For high SNR values, both Bayesian methods attain similar values of MSE, but the proposed empirical Bayes methodology is five times faster. We want to point out that these general conclusions depend a lot on the parameters used for the solver of the MAP estimation problem (in this case SALSA [1]). We included a detailed analysis of this in Appendix C.

5.3 Hyperspectral Unmixing with TV-SUnSAL

Hyperspectral sensors acquire hundreds of narrow band spectral images in different frequency bands. These images are collected in a three-dimensional hyperspectral data cube for processing and analysis. Although the spectral resolution is high, the spatial resolution is usually low, leading to the existence of "mixed" spectra in the acquired image pixels [44]. Hyperspectral unmixing is a source separation problem that aims at decomposing each mixed pixel into its constituent spectra (the so-called end-members) and their corresponding fractional abundances or proportions. This is normally done under the assumption of a linear mixing model [79]. In particular, linear unmixing techniques assume the availability of a library of spectral signatures and use the following model:

$$y = \mathrm{A}x + \mathrm{w}$$

where $y \in \mathbb{R}^{d_{\mathtt{f}} \times d_{\mathtt{p}}}$ is the hyperspectral image with $d_{\mathtt{f}}$ frequency channels and $d_{\mathtt{p}}$ pixels; $x \in \mathbb{R}^{d_{\mathtt{m}} \times d_{\mathtt{p}}}$ is the fractional abundance matrix compatible with the library $\mathrm{A} \in \mathbb{R}^{d_{\mathtt{f}} \times d_{\mathtt{m}}}$ containing the pure spectral signatures for $d_{\mathtt{m}}$ different materials; and w is a $d_{\mathtt{f}} \times d_{\mathtt{p}}$ Gaussian random variables with zero mean and covariance matrix $\sigma\,\mathrm{Id}$ and $\sigma > 0$. In [44], the unmixing problem is solved by using the regulariser g given for any $x \in \mathbb{R}^{d_{\mathtt{m}}} \times \mathbb{R}^{d_{\mathtt{p}}}$ by

$$g(x) = (\mathrm{TV}(x), \|x\|_1) \quad \text{s.t.} \quad x \geqslant 0\,,$$

which is associated with a two dimensional regularisation parameter $\theta = (\theta^{\mathrm{TV}}, \theta^1) \in \mathbb{R}^2$. $\theta^{\mathrm{TV}} \in \mathbb{R}$ controls the spatial cohesion of the objects, and $\theta^1 \in \mathbb{R}$ enforces sparsity on x. In this experiment, TV is the vectorial isotropic total variation pseudo-norm given for any $x \in \mathbb{R}^{d_{\mathtt{m}}} \times \mathbb{R}^{d_{\mathtt{p}}}$ by

$$\mathrm{TV}(x) = \sum_{i=1}^{d_{\mathtt{p}}} \sum_{j \in \mathcal{V}_i} \|x_i - x_j\|_1\ ,$$

where for any $i \in \{1, \ldots, d_{\mathtt{p}}\}$, $x_i \in \mathbb{R}^{d_{\mathtt{m}}}$ denotes the i-th image pixel and $\mathcal{V}_i$ its vertical and horizontal neighbour pixels.

Although this regulariser is not separable and we would therefore have to use Algorithm 3 with two MCMC chains, we can use a pseudo-likelihood approximation estimate θ using a single MCMC chain together with the expression of $\nabla_\theta \log \mathrm{Z}(\theta)$ for the homogeneous case. In this way we can achieve highly competitive computing times as well as compare our results with the hierarchical Bayesian method from [64], which we would otherwise not be able to apply to this problem.

More precisely, we consider $[\partial \log \mathrm{Z}/\partial\theta^1](\theta) = d/\theta^1$ and $[\partial \log \mathrm{Z}/\partial\theta^{TV}](\theta) = d/\theta^{TV}$. Although $x \mapsto \mathrm{TV}(x)$ and $x \mapsto \|x\|_1$ are not acting on independent subsets of x, we have empirically observed that this provides a good approximation and delivers excellent results.

We consider the experiment A-*Simulated Data Sets* case 1) *Simulated Data Cube 1* presented in [44, Section 4], particularly the case where w is a white Gaussian noise. In this experiment a synthetic hyperspectral image is generated by using five randomly selected spectral signatures. The image has $d_{\mathtt{p}} = 75 \times 75 = 5625$ pixels and $d_{\mathtt{f}} = 224$ frequency bands per pixel. For full details see [44]. We follow the exact same procedure as presented there, except for a modification in the spectral signature dictionary A. In [44] they consider a dictionary $\mathrm{A} \in \mathbb{R}^{224 \times 240}$, which is a library generated from a random selection of 240 materials from the USGS library[2]. Here we consider a simplified version where we only select $d_{\mathtt{m}} = 12$ random materials, thus having $\mathrm{A} \in \mathbb{R}^{224 \times 12}$. Out of these 12 materials, only 5 are present in the synthetic image. The synthetic fractional abundances x^0 are displayed in the first row of Figure 11 (only the 5 present end-members are shown)

We use the proposed algorithm to estimate θ^{TV} and θ^1 for this setup using Algorithm 2 under three different noise levels: we consider a SNR of 20 dB, 30 dB and 40 dB. For comparison, we also report the results obtained with the joint MAP method from [64] and by using the oracle value $\theta^\dagger$ that maximises the estimation signal-to-reconstruction-error (SRE) given by $\|x^0\|_2^2/\|x^0 - \hat{x}_{\mathrm{MAP}}\|_2^2$.

[2] Available online: http://speclab.cr.usgs.gov/spectral.lib06

We evaluated the proximal operator of $x \mapsto \theta^{\mathrm{TV}}\mathrm{TV}(x) + \theta^1 \|x\|_1$ using SUnSAL solver from [44] with 20 iterations. We address the positivity constraint separately by using its Moreau-Yosida envelope (14), leading to the additional term $x \mapsto (x - \Pi_+(x))/\lambda$ where Π_+ is the projection operator onto $[0, +\infty)^{d_{\mathtt{m}}} \times [0, +\infty)^{d_{\mathtt{p}}}$, and λ is the same smoothing parameter used for the other proximal operators.

To speed up the convergence, we use a gradient preconditioning technique explained in Appendix B.4. Since we use the preconditioned gradient of f_y instead of the gradient of f_y, the Lipschitz constant becomes $\mathtt{L} = 1/\sigma^2$. The algorithm parameters are chosen following the recommendations provided in Appendix B.1; we set $\theta_0^1 = 10$, $\theta_0^{\mathrm{TV}} = 10$, we initialised X_0 using the pseudo-inverse of A and projecting on the space of positive matrices. In addition, we perform 200 warm-up iterations and set for any $n \in \mathbb{N}^*$, $\delta_n = n^{-0.8}/(d_{\mathtt{p}} d_{\mathtt{m}})$.

Special care was taken when setting $\gamma > 0$ and $\lambda > 0$ due to the preconditioning. We set $\gamma_n = 1/(\mathtt{L} + 2/\lambda)$ for any $n \in \mathbb{N}$ and $\lambda = 0.9 \times \lambda_{\mathrm{A}}/\mathtt{L}$, where λ_{A} is the largest eigenvalue of $(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}$. We run the algorithm for 50 iterations and compute $(\bar{\theta})_{N \in \mathbb{N}}$ as defined in (28) with $N_0 = 30$.

In Figure 11 we display the MAP recovery of the synthetic fractional abundances using the estimated values of θ^{TV} and θ^1 with the SunSAL solver for SNR=30 dB.

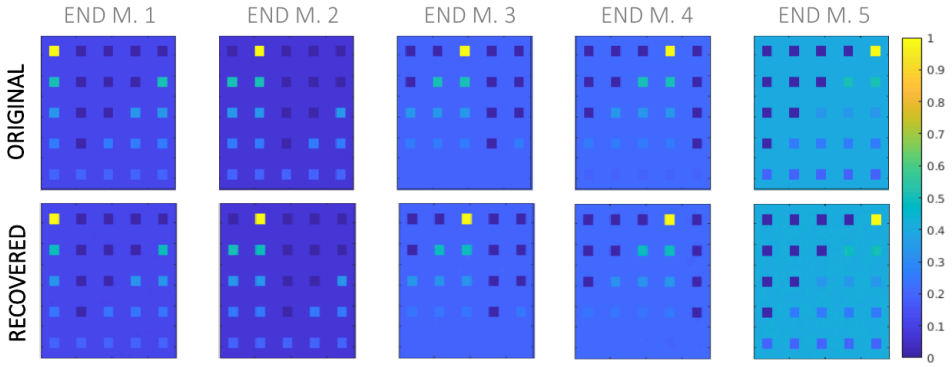


Figure 11: Hyperspectral Unmixing - Synthetic fractional abundances for 5 end-members. Original and MAP estimates for SNR=30 dB using the empirical Bayes posterior (7).

Figure 12 shows the evolution of the iterates $(\theta_n^1)_{n \in \mathbb{N}}$ and $(\theta_n^{\mathrm{TV}})_{n \in \mathbb{N}}$ and the relative change in the running averages $(|\bar{\theta}_{N+1} - \bar{\theta}_N|/\bar{\theta}_N)_{N \in \mathbb{N}}$ throughout iterations for SNR=30 dB. Observe the excellent convergence properties of the proposed scheme, which stabilises in as little as 25 iterations.

The obtained results are reported in Table 3 and summarised in Figure 13, which shows the signal to reconstruction error (SRE) surfaces for different values of the regularisation parameters. Observe that the empirical Bayesian method yields good results for all SNR values, and clearly outperforms the hierarchical Bayesian method for low SNR values. For high SNR values the hierarchical method achieved slightly better results. As discussed in section Section 3.5, we believe that this is due to the fact that, at high SNR values, the likelihood $x \mapsto p(y|x)$ dominates the posterior and mitigates errors related to the misspecification of the prior. More precisely, if the hyperprior that we set on θ assigns a high weight to values of θ that lead to bad models, *i.e.* a misspecified prior $x \mapsto p(x|\theta)$, the

Method		SNR=20 dB		SNR=30 dB		SNR=40 dB	
	Stop criteria	SRE	Time (s)	SRE	Time (s)	SRE	Time (s)
$\theta_\dagger$ (Oracle)	–	29.38	–	38.61	–	47.64	–
E.B.	50 iters.	27.46	36	38.42	37	45.68	42
H.B. [64]	15 iters.	18.33	76	31.72	77	47.36	76

Table 3: Hyperspectral unmixing - Signal to reconstruction error (SRE) obtained for different SNR values along with computing times expressed in seconds.

impact of this misspecification on the recovered estimates depends on the degree of concentration of

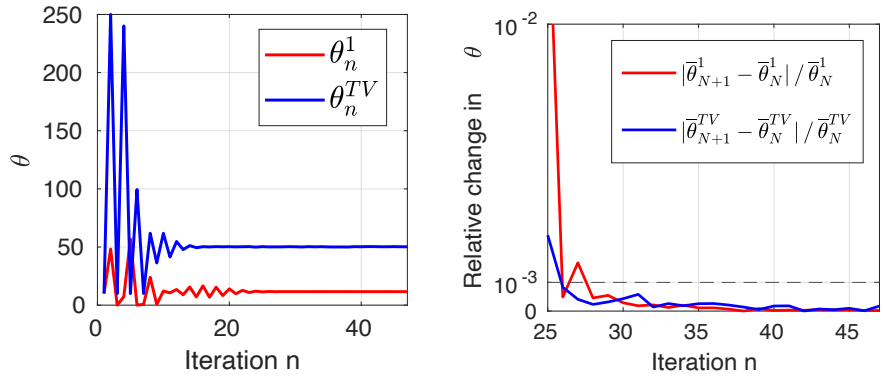
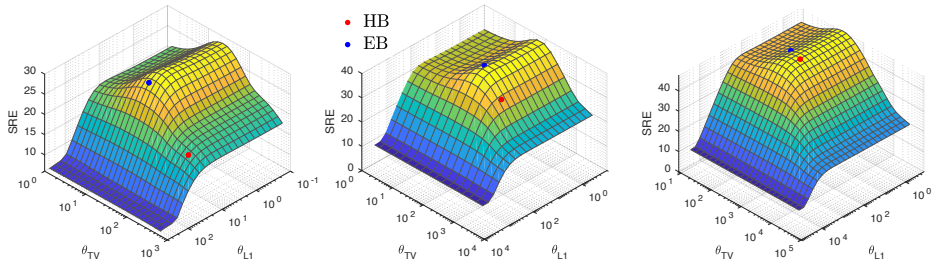


Figure 12: Hyperspectral Unmixing - Evolution of the iterates $(\theta_n^1)_{n\in\mathbb{N}}$ and $(\theta_n^{\mathrm{TV}})_{n\in\mathbb{N}}$ (left) and of the relative successive differences $(|\bar{\theta}_{N+1} - \bar{\theta}_N|/\bar{\theta}_N)_{N\in\mathbb{N}}$ (right) for the proposed method with SNR=30 dB. The relative change is computed after 25 burn-in iterations.



(a) SNR=20 dB (b) SNR=30 dB (c) SNR=40 dB

Figure 13: Hyperspectral Unmixing - Signal to reconstruction error (SRE) surfaces for different SNR values expressed in dB. Comparison between parameters estimated with our empirical Bayesian algorithm (EB) and with the hierarchical Bayesian method (HB) from [64].

the likelihood. At high SNR, the likelihood dominates the posterior thus concealing the possible prior misspecification and leading to good results. Conversely, at low SNR values, the performance of the hierarchical model is degraded by model misspecification.

Also note in Table 3 that the computing times for the empirical Bayesian method are approximately two times faster than the ones for the hierarchical method.

5.4 Denoising with a total generalized variation prior

In this last experiment, we apply the proposed methodology to a challenging problem that is beyond the scope of the considered class of models and our theoretical guarantees. We consider an image denoising problem where $y \sim \mathcal{N}(x, \sigma^2 \mathrm{Id})$ with $\sigma^2 > 0$ and where we use the following prior

$$p(x|\theta^1, \theta^2) = \frac{1}{\mathrm{Z}(\theta^1, \theta^2)} \exp\{-\mathrm{TGV}^2_{\theta^1,\theta^2}(x) - \varepsilon\|x\|_2^2\} \,,$$

where $\varepsilon > 0$ and where $\mathrm{TGV}^2_{\theta^1,\theta^2}(x)$ is a second-order generalisation of the conventional total variation regulariser, given, for any $(\theta^1, \theta^2) \in [0, +\infty)^2$ and $x \in \mathbb{R}^d$, by

$$\mathrm{TGV}^2_{\theta^1,\theta^2}(x) = \min_{r\in\mathbb{R}^{2d}} \{\theta^1 \|r\|_{1,2} + \theta^2 \|J(\Delta x - r)\|_{1,\mathrm{Frob.}}\} \,. \tag{29}$$

where $\Delta = (\Delta^v, \Delta^h)$ is the discrete image-gradient operator that computes the first-order vertical and horizontal pixel differences, and J computes the Jacobian matrix of the image-gradient vector

field to capture second-order information (i.e., $(J\Delta)(x)$ is a discrete image-Hessian operator) [25]. This generalisation was first considered in [19] and further studied in [16] as a means of incorporating second-order derivative information to eliminate the common staircasing artifacts associated with the conventional TV regulariser.

A main difficulty associated with using the TGV regulariser is the need to correctly set the parameters θ^1 and θ^2, which control the strength as well as the characteristics of the regularisation enforced (as explained in [25], the TGV regularisation behaves like the standard TV regularisation for large θ^2 values, whereas for small values it behaves like the ℓ_1-Frobenius norm of the discrete image-Hessian). Figure 14 below illustrates the dramatic effect that these two parameters have on the quality of the recovered MAP estimate. Observe the strong coupling between θ^1 and θ^2, which makes setting their values particularly challenging.

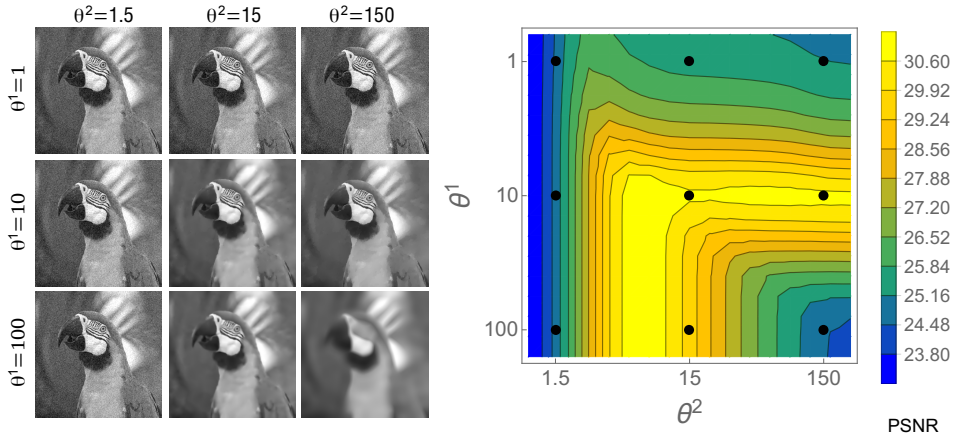


Figure 14: Denoising with TGV prior. MAP estimates for different values of θ^1 and θ^2 for `parrot` image with SNR= 5.6 dB (left). PSNR for different values of θ^1 and θ^2 (right). The 9 black points on the right plot show the location of the parameter combinations used to compute the MAP estimates on the left.

However, this prior is not in the exponential family because θ^1 and θ^2 play a role in the definition of the statistic $\mathrm{TGV}^2_{\theta^1,\theta^2}(x)$. Therefore, our methodology and theory do not directly apply. Also note that the additional regularisation $\varepsilon\|x\|_2^2$ with $\varepsilon > 0$ is necessary to guarantee that $p(x)$ is proper, which is potentially important in order to apply the proposed methodology with two Markov chains (otherwise the auxiliary chain targeting $p(x)$ would not be ergodic - two chains are required because (29) is not separable and homogeneous). We use $\varepsilon = 10^{-10}$.

In order to apply the proposed methodology to the estimation of θ^1 and θ^2 we use an approximation of the gradient $\nabla_\theta \log p(x|\theta^1,\theta^2)$. More precisely, we express $p(x)$ as follows for any $x \in \mathbb{R}^d$ and $\theta^1,\theta^2 > 0$

$$p(x|\theta^1,\theta^2) = \frac{1}{\mathrm{Z}(\theta^1,\theta^2)} \exp\left[-\theta^1 g_1(x,\theta^1,\theta^2) - \theta^2 g_2(x,\theta^1,\theta^2) - \varepsilon\|x\|_2^2\right] ,$$

with

$$g_1(x,\theta^1,\theta^2) = \left\|r(x,\theta^1,\theta^2)\right\|_{1,2} ,$$

$$g_2(x,\theta^1,\theta^2) = \left\|J(\Delta x - r(x,\theta^1,\theta^2))\right\|_{1,\mathrm{Frob.}} ,$$

$$r(x,\theta^1,\theta^2) = \operatorname*{argmin}_{s\in\mathbb{R}^{2d}} \{\theta^1 \left\|s\right\|_{1,2} + \theta^2 \left\|J(\Delta x - s)\right\|_{1,\mathrm{Frob.}}\} ,$$

and approximate the partial derivatives $\frac{\partial}{\partial\theta^1}\log p(x|\theta^1,\theta^2)$ and $\frac{\partial}{\partial\theta^2}\log p(x|\theta^1,\theta^2)$ by

$$\frac{\partial}{\partial\theta^1}\log p(x|\theta^1,\theta^2) \approx \mathrm{E}_{x|\theta^1,\theta^2}[g_1(x,\theta^1,\theta^2)] - g_1(x,\theta^1,\theta^2) ,$$

$$\frac{\partial}{\partial\theta^2}\log p(x|\theta^1,\theta^2) \approx \mathrm{E}_{x|\theta^1,\theta^2}[g_2(x,\theta^1,\theta^2)] - g_2(x,\theta^1,\theta^2) .$$

This approximation of the gradient, which arises from omitting the terms

$$\mathrm{E}_{x|\theta^1,\theta^2}\left[\theta^1\frac{\partial}{\partial\theta^1}g_1(x,\theta^1,\theta^2)+\theta^2\frac{\partial}{\partial\theta^1}g_2(x,\theta^1,\theta^2)\right]-\theta^1\frac{\partial}{\partial\theta^1}g_1(x,\theta^1,\theta^2)-\theta^2\frac{\partial}{\partial\theta^1}g_2(x,\theta^1,\theta^2)$$

and

$$\mathrm{E}_{x|\theta^1,\theta^2}\left[\theta^1\frac{\partial}{\partial\theta^2}g_1(x,\theta^1,\theta^2)+\theta^2\frac{\partial}{\partial\theta^2}g_2(x,\theta^1,\theta^2)\right]-\theta^1\frac{\partial}{\partial\theta^2}g_1(x,\theta^1,\theta^2)-\theta^2\frac{\partial}{\partial\theta^2}g_2(x,\theta^1,\theta^2)$$

in the calculation of the partial derivatives $\frac{\partial}{\partial\theta^1}\log p(x|\theta^1,\theta^2)$ and $\frac{\partial}{\partial\theta^2}\log p(x|\theta^1,\theta^2)$, introduces an additional bias in the stochastic gradients driving Algorithm 3[3]. However, the numerical experiments reported below suggest that the algorithm is robust to this additional bias, in the sense that we empirically observe good convergence to useful estimates of θ^1 and θ^2.

In our experiments, we implement Algorithm 3 with this approximate gradient and follow the recommendations provided in Appendix B.1 to set the algorithm parameters; we perform 25 warm-up iterations and set $\theta_0^1=\theta_0^2=10$, $X_0=X_0=y$, for any $n\in\mathbb{N}^*$, $\delta_n=20\times n^{-0.8}/d$, and we set $\lambda=\min\left(5L^{-1},\lambda_{\max}\right)$ with $\lambda_{\max}=2$ and $L=(0.95/\sigma)^2$. To stop the algorithm we consider three different cases: we stop the algorithm i) after $N=2000$ fixed iterations ii) when the relative change in $\bar{\theta}_N$ is $\|\bar{\theta}_{N+1}-\bar{\theta}_N\|_\infty\leqslant 10^{-4}$ and iii) $\|\bar{\theta}_{N+1}-\bar{\theta}_N\|_\infty\leqslant 10^{-3}$. Again, we skip the first 20 iterations before computing the average sequence $(\bar{\theta}_N)_{N\in\mathbb{N}}$, *i.e.* $(\bar{\theta}_N)_{N\in\mathbb{N}}$ is given by (28) with $N_0=20$. We also considered a thinning of 6 iterations in the chain associated with the prior as its samples were roughly 6 times more correlated than those coming from the chain targeting the posterior (i.e., we discard 5 every 6 samples as explained in Appendix B.3). To compute the $\mathrm{TGV}^2_{\theta^1,\theta^2}$ norm and proximal operator, we use the iterative primal-dual algorithm [25].

Applying Algorithm 3 to the entire image is too computationally expensive because of the complexity associated with evaluating the proximal operator of the TGV regulariser. Therefore, in this experiment we estimate $\bar{\theta}_N$ from a representative patch of size 255×255 pixels, and then use the estimated θ^1 and θ^2 values to compute the MAP estimate of the entire image[4]. We consider the same ten test images used in Section 5.2 and we set the noise variance σ^2, such that the signal-to-noise-ratio (SNR) is 8 dB, 12 dB, or 20 dB. For each image and noise level, we first obtain an estimate for θ^1 and θ^2 and then use them to compute the MAP estimator $\hat{x}_{\mathrm{MAP}}$ (given by (4)) using the same solver [25] we use for the proximal operator. We measure estimation performance by computing the peak-signal-to-noise-ratio (PSNR) given by $\mathrm{PSNR}(x,\hat{x}_{\mathrm{MAP}})=-10\log_{10}\|x-\hat{x}_{\mathrm{MAP}}\|_2^2/d$. All the PSNR plots shown in Figure 16, Figure 17 and Figure 20 were computed with the entire image.

Table 4 below summarises the average PSNR values and average computing times obtained for each SNR value for the three different stopping criteria. We observe that the proposed empirical Bayesian method achieves very good results for all SNR values and is very close to the oracle performance. Crucially, the stopping criteria has a strong impact on the computing times but not on the resulting PSNR values. Therefore, although convergence can take close to one hour with a strict convergence criterion, good results can be obtained in the order of a minute by using a weaker convergence criterion.

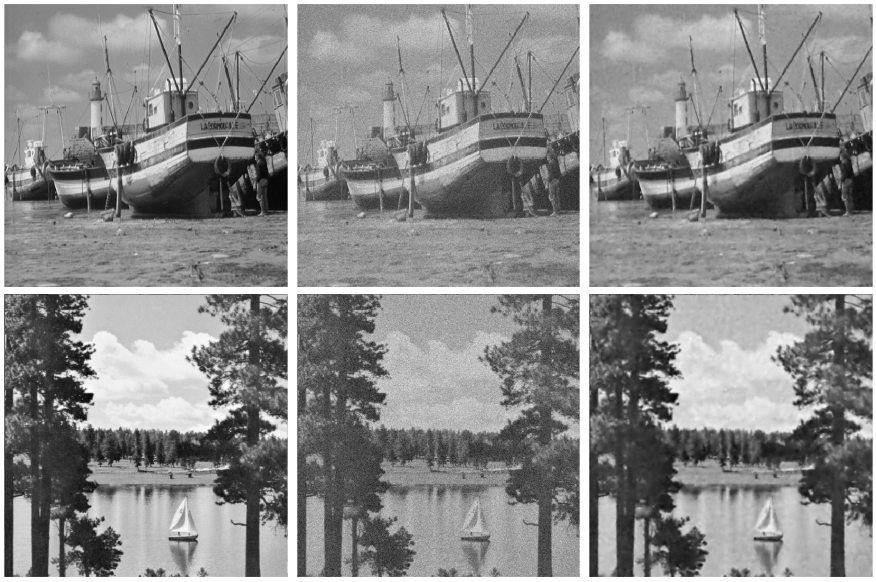
For illustration, Figure 15 depicts the original image, the noisy observation and the recovered MAP estimates for the `boat` and `lake` test images with SNR = 8 dB.

More interestingly, Figure 16 and Figure 17 show the landscape of the PSNR as a function of θ^1 and θ^2 for the two test images, with the obtained solutions highlighted as a blue dot. Observe that the estimated solutions are extremely close to the optimal ones, which is remarkable given the difficulty of the problem and the fact that solutions are derived directly from statistical inference principles, without any form of ground truth.

Following on from this, Figure 18 and Figure 19 show respectively the evolution of the iterates and the relative change in the estimated values of θ^1 and θ^2, for the `lake` test image, and for SNR = 8 dB, = 12 dB, and = 20 dB. Observe that the algorithm converges very quickly and can deliver a useful solution in approximately 50 iterations if the weaker convergence criterion is used, or in approximately 500 iterations if one uses a stricter convergence criterion.

[3] A rigorous analysis of this bias should also consider the points where $\mathrm{TGV}^2_{\theta^1,\theta^2}(x)$ is not differentiable w.r.t. θ^1 and θ^2. This can be achieved by using similar techniques to [80].

[4] For homogeneous regularisers, θ is asymptotically independent of the dimension of x when d is large, suggesting that it is possible to estimate its value from a representative image patch. Our empirical results suggest that this might hold for other models as well.



(a) Original (b) Degraded (c) Empirical Bayes

Figure 15: Denoising with TGV prior for `boat` and `lake` test images: (a) True image, (b) noisy observation y (SNR=8 dB), (c) MAPestimators obtained with empirical Bayes.

Lastly, Figure 20 below explores the robustness to different initialisations by showing the evolution of the iterates on the landscape of PSNR values for the `flinstones` image with SNR = 12 dB. We consider three different initialisations, highlighted in colours red, green, and blue, and observe that in the three cases the algorithm quickly converges to values for the parameters θ^1 and θ^2 that are close-to-optimal in terms of the resulting PSNR. However, the algorithm is not fully robust to bad initialisation because of the non-convexity and the approximations involved. For example, initialising the algorithm in the corner of the PSNR landscape (e.g., $\theta_0^1 = \theta_0^2 = 100$) does not lead to a satisfactory solution, indicating that a careful initialisation is required. Alternatively, one could also initialise the algorithm by performing a certain number of updates on θ^1 with θ^2 fixed to a small value - e.g. $\theta^2 = 1$ - to keep the model close to the conventional total variation regulariser, and then update both θ^1 with θ^2 until the convergence criterion is satisfied.

To conclude, we note that there are several other generalisations of the total variation regularisation (see [16]). We have chosen to perform our experiments with (29) because of the availability of the efficient MATLAB implementation [25]. However, we expect that Algorithm 3 will also perform well for other generalisations of the total variation norm, particularly the second-order generalisation proposed in [16] that is very similar to (29).

Method	SNR=8 dB		SNR=12 dB		SNR=20 dB	
	PSNR	Time	PSNR	Time	PSNR	Time (min)
$\theta_\dagger$ (Oracle)	27.8 ± 2.35		30.2 ± 2.12		35.6 ± 1.77	
$_{2000\ \text{iter}}$ E.B.	27.1 ± 2.77	38.90	29.7 ± 2.29	37.7	35.4 ± 1.80	38.6
$_{tol\ 10^{-4}}$ E.B.	27.0 ± 2.85	2.11	29.7 ± 2.31	6.24	35.5 ± 1.80	10.9
$_{tol\ 10^{-3}}$ E.B.	26.9 ± 2.97	0.21	29.7 ± 2.32	0.81	35.5 ± 1.82	1.28

Table 4: Denoising with TGV prior. Average mean squared error ± standard deviation obtained for ten different images. We show results for different stopping criteria, either with a fixed number of iterations or with a maximum tolerance for the relative change in the mean θ^1 and θ^2 estimates.

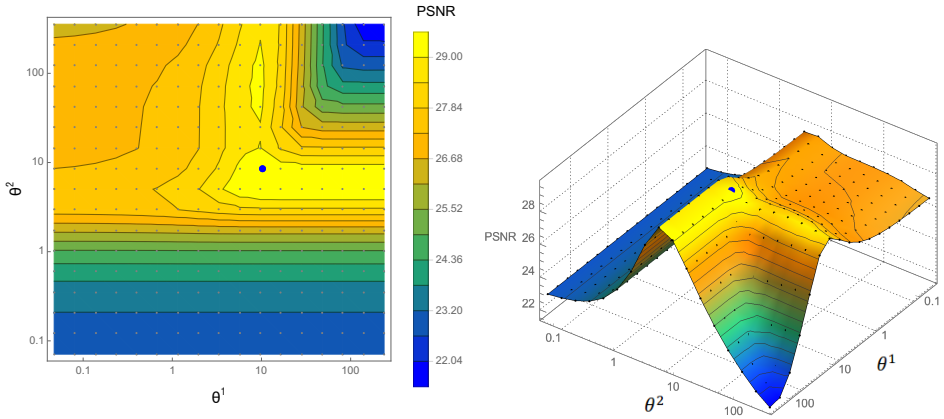


Figure 16: Denoising with TGV prior on `boat` image (SNR=8 dB). PSNR for different values of θ^1 and θ^2. Blue marker shows the location of $\bar{\theta}_N$ estimated with empirical Bayes using 2000 iterations.

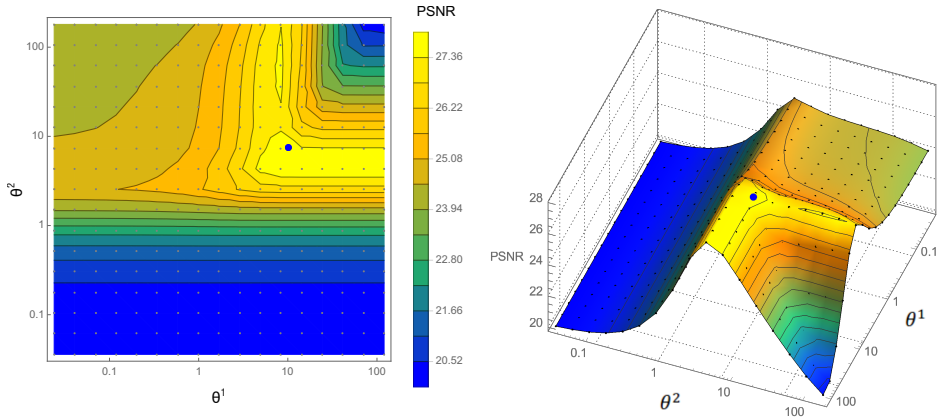
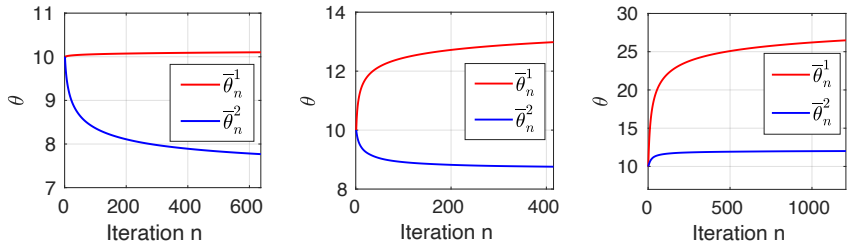
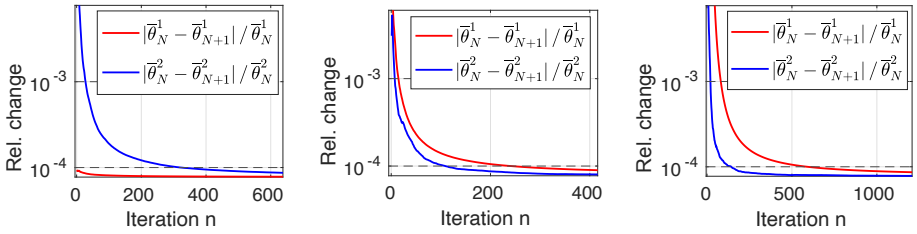


Figure 17: Denoising with TGV prior on `lake` image (SNR=8 dB). PSNR for different values of θ^1 and θ^2. Blue marker shows the location of $\bar{\theta}_N$ estimated with empirical Bayes using 2000 iterations.



(a) SNR 8dB (b) SNR 12dB (c) SNR 20dB

Figure 18: Denoising with TGV prior . Evolution of the iterates $(\theta_n^1)_{n\in\mathbb{N}}$ and $(\theta_n^2)_{n\in\mathbb{N}}$ for the `lake` test image for different SNR values.



(a) SNR 8dB (b) SNR 12dB (c) SNR 20dB

Figure 19: Denoising with TGV prior. Relative successive differences $|\bar{\theta}^i_N - \bar{\theta}^i_{N+1}|/\bar{\theta}^i_N$ with $i = 1, 2$ for the proposed method with the `lake` test image for different SNR values..

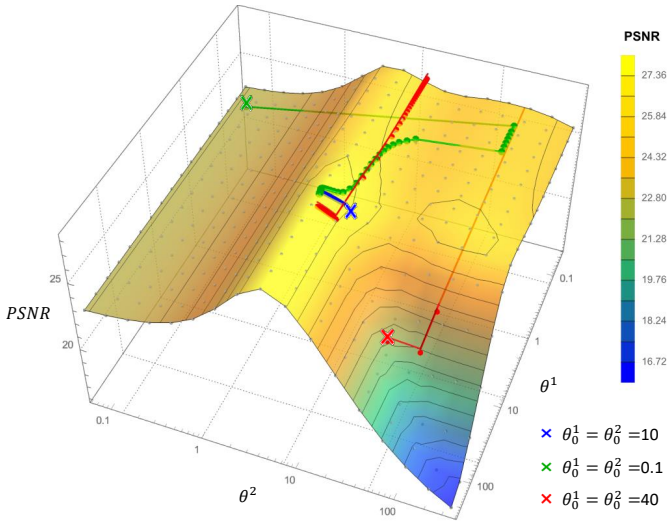


Figure 20: Denoising with TGV prior on the `flinstones` image (SNR=12 dB). Evolution of the iterates $(\theta^1_n)_{n\in\mathbb{N}}$ and $(\theta^2_n)_{n\in\mathbb{N}}$ for different initial values θ^1_0 and θ^2_0. When initialising with $\theta^1_0 = \theta^2_0 = 40$ (red) the algorithm converges to a different point with a similar PSNR.

6 Conclusions

This paper considered the automatic selection of regularisation parameters in imaging inverse problems, with a particular focus on problems that are convex w.r.t. the unknown image and possibly non-smooth, and which would be typically solved by maximum-a-posteriori estimation by using modern proximal optimisation techniques. We adopted an empirical Bayesian approach and proposed a computational method to efficiently and accurately estimate regularisation parameters by maximum marginal likelihood estimation. The considered marginal likelihood function is computationally intractable and we proposed to address this difficulty by using a stochastic proximal gradient optimisation algorithm that is driven by two proximal MCMC samplers, and which tightly combines the strengths of modern high-dimensional optimisation and Monte Carlo sampling techniques. Because the proposed method uses the same basic operators as proximal optimisation algorithms, namely gradient and proximal operators, it is straightforward to apply to problems that are currently solved by proximal optimisation. Moreover, we provided a detailed theoretical analysis of the proposed methodology, including easily verifiable conditions for convergence. In addition to being highly computational efficient and having strong theoretical underpinning, the proposed methodology is very general and can be used to simultaneously estimate multiple regularisation parameters, unlike some alternative approaches from the

literature that can only handle a single or scalar parameter.

We demonstrated the methodology with a range of imaging problems and models. We first considered image denoising and non-blind deblurring problems involving scalar regularisation parameters, and showed that the method achieved close-to-optimal performance in terms of MSE and outperformed alternative approaches from the literature. We then successfully applied the method to two challenging problems involving bivariate regularisation parameters: a sparse hyperspectral unmixing problem with a total-variation plus sparsity prior, and a challenging denoising problem using a second-order total generalised variation regulariser. Again, the method delivered close-to-optimal results, as measured by estimation MSE.

Future work will focus on relaxing the convexity assumptions to provide theoretical convergence guarantees for non-convex problems, and on improving computational efficiency by using the recently proposed accelerated proximal Markov kernels [84]. The application of the proposed methodology to challenging problems arising in medical and astronomical imaging is currently under investigation. Another important perspective for future work is to extend this methodology to semi-blind and blind imaging problems.

7 Acknowledgements

We are grateful to Dr. Charles Deledalle for providing us with a SUGAR implementation for an ADMM solver available at `https://github.com/deledalle/sugar/blob/master/solvers/admm.m`. AD acknowledges financial support from Polish National Science Center grant: NCN UMO-2018/31/B/ST1/00253.

References

[1] Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(9):2345–2356, 2010.

[2] Mariana SC Almeida and Mário AT Figueiredo. Parameter estimation for blind and non-blind deblurring using residual whiteness measures. *IEEE Transactions on Image Processing*, 22(7):2751–2763, 2013.

[3] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[4] Yves F Atchade. An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability*, 8(2):235–254, 2006.

[5] Yves F. Atchadé. A computational framework for empirical Bayes inference. *Statistics and Computing*, 21(4):463–473, 2011.

[6] Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res*, 18(1):310–342, 2017.

[7] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. Parameter estimation in TV image restoration using variational distribution approximation. *IEEE transactions on image processing*, 17(3):326–339, 2008.

[8] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. Variational Bayesian super resolution. *IEEE Transactions on Image Processing*, 20(4):984–999, 2011.

[9] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

[10] Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab.*, 13:60–66, 2008.

[11] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017. With a foreword by Hédy Attouch.

[12] Federico Benvenuto and Cristina Campi. A discrepancy principle for the landweber iteration based on risk minimization. *Applied Mathematics Letters*, 96:1 – 6, 2019.

[13] Sebastian Berisha, James G Nagy, and Robert J Plemmons. Deblurring and sparse unmixing of hyperspectral images using multiple point spread functions. *SIAM Journal on Scientific Computing*, 37(5):S389–S406, 2015.

[14] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379, 2012.

[15] Sergey Bobkov and Mokshay Madiman. Concentration of the information in data with log-concave distributions. *Annals of Probability*, 39(4):1528–1543, 2011.

[16] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[17] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[18] Bradley P. Carlin and Thomas A. Louis. Empirical Bayes: past, present and future. *J. Amer. Statist. Assoc.*, 95(452):1286–1289, 2000.

[19] Antonin Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76:167–188, 1997.

[20] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

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

[22] Emilie Chouzenoux, Anna Jezierska, Jean-Christophe Pesquet, and Hugues Talbot. A Convex Approach for Image Restoration with Exact Poisson–Gaussian Likelihood. *SIAM Journal on Imaging Sciences*, 8(4):2662–2682, 2015.

[23] Julianne Chung and Linh Nguyen. Motion estimation and correction in photoacoustic tomographic reconstruction. *SIAM Journal on Imaging Sciences*, 10(1):216–242, 2017.

[24] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[25] Laurent Condat. Matlab code for total generalized variation denoising, 2016.

[26] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal. Efficient stochastic optimisation by unadjusted langevin monte carlo. application to maximum marginal likelihood and empirical bayesian estimation. *arXiv preprint arXiv:1906.12281*, 2019.

[27] Charles-Alban Deledalle, Samuel Vaiter, Jalal Fadili, and Gabriel Peyré. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487, 2014.

[28] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[29] Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series: Theory, methods and applications with R examples*. Chapman and Hall/CRC, 2014.

[30] Paul Dupuis and Richard S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication.

[31] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[32] Alain Durmus, Eric Moulines, et al. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[33] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[34] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[35] Yonina C Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.

[36] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[37] Gersende Fort, Edouard Ollier, and Adeline Samson. Stochastic Proximal Gradient Algorithms for Penalized Mixed Models. *arXiv preprint arXiv:1704.08891*, 2017.

[38] Bruno Galerne and Arthur Leclaire. Texture inpainting using efficient Gaussian conditional simulation. *SIAM Journal on Imaging Sciences*, 10(3):1446–1474, 2017.

[39] Raja Giryes, Michael Elad, and Yonina C Eldar. The projected GSURE for automatic parameter tuning in iterative shrinkage methods. *Applied and Computational Harmonic Analysis*, 30(3):407–422, 2011.

[40] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[41] Peter J Green, Krzysztof Latuszyński, Marcelo Pereyra, and Christian P Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862, 2015.

[42] Per Christian Hansen and Dianne Prost O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, 1993.

[43] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second edition, 1989.

[44] Marian-Daniel Iordache, José M Bioucas-Dias, and Antonio Plaza. Total variation spatial regularization for sparse hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4484–4502, 2012.

[45] Mark A Iwen, Aditya Viswanathan, and Yang Wang. Fast phase retrieval from local correlation measurements. *SIAM Journal on Imaging Sciences*, 9(4):1655–1688, 2016.

[46] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[47] Belhal Karimi, Blazej Miasojedow, Éric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. *arXiv preprint arXiv:1902.00629*, 2019.

[48] Michael Kech and Felix Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

[49] BT Knapik, BT Szabó, AW van der Vaart, and JH van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3-4):771–813, 2016.

[50] Solomon Kullback. *Information theory and statistics*. John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London, 1959.

[51] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 15. Siam, 1995.

[52] M. Lebrun, A. Buades, and J. Morel. A nonlocal bayesian image denoising algorithm. *SIAM Journal on Imaging Sciences*, 6(3):1665–1688, 2013.

[53] Shutao Li, Xudong Kang, Leyuan Fang, Jianwen Hu, and Haitao Yin. Pixel-level image fusion: A survey of the state of the art. *Information Fusion*, 33:100–112, 2017.

[54] Robert S. Liptser and Albert N. Shiryaev. *Statistics of random processes. II*, volume 6 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, expanded edition, 2001. Applications, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability.

[55] Felix Lucka, Katharina Proksch, Christoph Brune, Nicolai Bissantz, Martin Burger, Holger Dette, and Frank Wübbeling. Risk estimators for choosing regularization parameters in ill-posed problems-properties and limitations. *Inverse Problems & Imaging*, 12(5):1121–1155, 2018.

[56] Yosra Marnissi, Yuling Zheng, Emilie Chouzenoux, and Jean-Christophe Pesquet. A Variational Bayesian Approach for Image Restoration?Application to Image Deblurring With Poisson–Gaussian Noise. *IEEE Transactions on Computational Imaging*, 3(4):722–737, 2017.

[57] Rafael Molina, Aggelos K Katsaggelos, and Javier Mateos. Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Transactions on Image Processing*, 8(2):231–246, 1999.

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

[59] Veniamin I Morgenshtern and Emmanuel J Candes. Super-resolution of positive sources: The discrete setup. *SIAM Journal on Imaging Sciences*, 9(1):412–444, 2016.

[60] Vladimir Alekseevich Morozov. *Methods for solving incorrectly posed problems*. Springer Science & Business Media, 2012.

[61] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[62] Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

[63] Marcelo Pereyra. Revisiting Maximum-A-Posteriori Estimation in Log-Concave Models. *SIAM Journal on Imaging Sciences*, 12(1):650–670, 2019.

[64] Marcelo Pereyra, José M Bioucas-Dias, and Mário AT Figueiredo. Maximum-a-posteriori estimation with unknown regularisation parameters. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 230–234. IEEE, 2015.

[65] Marcelo Pereyra, Nicolas Dobigeon, Hadj Batatia, and Jean-Yves Tourneret. Estimating the granularity coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm. *IEEE Transactions on Image Processing*, 22(6):2385–2397, 2013.

[66] Marcelo Pereyra, Nicolas Dobigeon, Hadj Batatia, and Jean-Yves Tourneret. Computing the Cramer–Rao bound of Markov random field parameters: application to the Ising and the Potts models. *IEEE Signal Processing Letters*, 21(1):47–50, 2014.

[67] Marcelo Pereyra, Philip Schniter, Emilie Chouzenoux, Jean-Christophe Pesquet, Jean-Yves Tourneret, Alfred O Hero, and Steve McLaughlin. A survey of stochastic simulation and optimization methods in signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):224–241, 2016.

[68] Jean-Christophe Pesquet, Amel Benazza-Benyahia, and Caroline Chaux. A SURE approach for digital signal/image deconvolution problems. *IEEE Transactions on Signal Processing*, 57(12):4616–4632, 2009.

[69] S. Petrone, J. Rousseau, and C. Scricciolo. Bayes and empirical Bayes: do they merge? *Biometrika*, 101(2):285–302, 2014.

[70] George Pólya and Gabor Szegő. *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.

[71] Saiprasad Ravishankar and Yoram Bresler. Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging. *SIAM Journal on Imaging Sciences*, 8(4):2519–2557, 2015.

[72] H. Robbins. An empirical Bayes approach to statistics. In *Herbert Robbins Selected Papers*, pages 41–47. Springer, 1985.

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

[74] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[75] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[76] J. Rousseau and B. Szabo. Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.*, 45(2):833–865, 2017.

[77] Toby Sanders, Rodrigo B Platte, and Robert D Skeel. Maximum evidence algorithms for automated parameter selection in regularized inverse problems. *arXiv preprint arXiv:1812.11449*, 2018.

[78] Carola-Bibiane Schönlieb. *Partial Differential Equation Methods for Image Inpainting*, volume 29. Cambridge University Press, 2015.

[79] Miguel Simões, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3373–3388, 2015.

[80] De Bortoli V., Durmus A., Pereyra M., and Vidal A. Supplementary to "....". 2018.

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

[82] Cao Van Chung, JC De los Reyes, and CB Schönlieb. Learning optimal spatially-dependent regularization parameters in total variation image denoising. *Inverse Problems*, 33(7):074005, 2017.

[83] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[84] Luis Vargas, Marcelo Pereyra, and Konstantinos C. Zygalakis. Accelerating proximal Markov chain Monte Carlo by using explicit stabilised methods. *arXiv e-prints*, page arXiv:1908.08845, Aug 2019.

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

[86] Halbert White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25, 1982.

[87] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[88] Marcelo VW Zibetti, Fermín SV Bazán, and Joceli Mayer. Determining the regularization parameters for super-resolution problems. *Signal Processing*, 88(12):2890–2901, 2008.

A Fisher's identity

Fisher's identity is a standard result in the probability literature (e.g. see [29, Proposition D.4]). We reproduce its proof here for completeness.

Proposition 3. *For any $\theta \in \Theta \in \mathbb{R}^{d_\Theta}$ and $\tilde{x} \in \mathbb{R}^d$, let $(x, y) \mapsto p(x, y|\theta)$ and $y \mapsto p(y|\tilde{x})$ be positive probability density functions on $\mathbb{R}^d \times \mathbb{R}^{d_y}$ and $\mathbb{R}^{d_y}$. Assume that for any $x \in \mathbb{R}^d$ and $\theta \in \mathrm{int}(\Theta)$, $\theta \mapsto p(y, x|\theta)$ is differentiable. In addition, assume that for any $y \in \mathbb{R}^{d_y}$ and $\theta \in \mathrm{int}(\Theta)$, there exist $\varepsilon > 0$ and $\tilde{g}$ such that for any $\tilde{\theta} \in \overline{\mathrm{B}}(\theta, \varepsilon)$ and $x \in \mathbb{R}^d$, $\|\nabla_\theta p(y, x|\tilde{\theta})\| \leqslant \tilde{g}(x)$ with $\int_{\mathbb{R}^d} \tilde{g}(x) p(y|x) \mathrm{d}x < +\infty$. Then, for any $y \in \mathbb{R}^{d_y}$, $\theta \mapsto p(y|\theta)$ is differentiable over $\mathrm{int}(\Theta)$ and we have for any $y \in \mathbb{R}^{d_y}$ and $\theta \in \mathrm{int}(\Theta)$,*

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

Proof. Let $y \in \mathbb{R}^{d_y}$. It is clear using the Leibniz integral rule that $\theta \mapsto p(y|\theta)$ is differentiable over $\mathrm{int}(\Theta)$ and we have for any $\theta \in \mathrm{int}(\Theta)$

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

which concludes the proof. □

B Practical implementation guidelines

In this section we provide some guidelines for the practitioner. These recommendations do not seek to define optimal values for specific models, but rather to provide general rules that are simple to use and robust. The section is organised as follows: Appendix B.1 discusses suitable ranges for the parameters appearing in Algorithm 1, Algorithm 2 and Algorithm 3, as well as some important considerations on the parameter space Θ. The impact of differences in convergence speed between the two chains of Algorithm 3 is discussed in Appendix B.3. Appendix B.4 discusses the impact of the convergence speed of MYULA on the convergence speed of the SAPG schemes. Lastly, Appendix B.5 studies the impact of the bias in MYULA on the estimates produced by the SAPG scheme.

B.1 Setting the algorithm parameters

Algorithm 1 depends on three parameters: the smoothing parameter $\lambda > 0$; the discretisation step-size $(\gamma_n)_{n \in \mathbb{N}}$; and the learning rate in the perturbed gradient descent $(\delta_n)_{n \in \mathbb{N}}$.

Selecting λ The parameter λ controls the regularity properties of g^λ. Although it would be natural to use the same smoothing for both MCMC kernels, this is not strictly necessary. In [33], it is recommended to use values of $\lambda = \mathcal{O}(\mathrm{L}_y^{-1})$ because there is no benefit in using larger values of λ as γ will still saturate at L_y^{-1}. However, for any $\theta \in \Theta$ and $\gamma > 0$ the kernel $\bar{\mathrm{R}}_{\gamma,\theta}$ is not affected by f_y, since it only targets the prior distribution. So in this case, picking a larger value of smoothing parameter λ' might be beneficial as it would enable using a larger step-size $\gamma' > 0$. As we will explain in Appendix B.3 and Appendix B.4 this might sometimes be a useful strategy for increasing the convergence speed of the algorithm. Therefore a reasonable criterion is to set $\lambda \in (\mathrm{L}_y^{-1}, 10\mathrm{L}_y^{-1})$ and $\lambda = \lambda'$ as long as L_y^{-1} is not too large (usually we try to keep $\lambda < 10$ to avoid introducing too much smoothing bias). In most of our experiments we set $\lambda = \min(5\mathrm{L}_y^{-1}, \lambda_{\max})$ with $\lambda_{\max} = 2$.

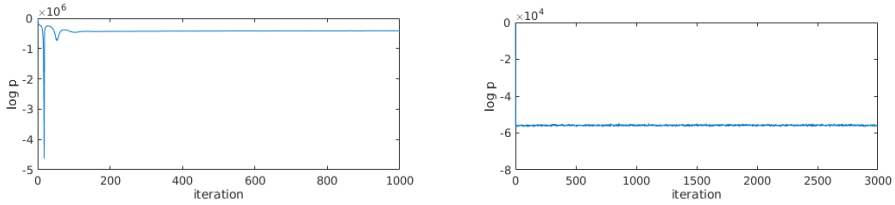
Selecting γ Because ∇f_y is Lipschitz continuous with constant $\mathrm{L}_y > 0$, we know from the remark following Theorem 2 that γ and γ' should take values in the range $(0, 2/(\mathrm{L}_y + \lambda^{-1}))$ and $(0, 2\lambda')$ respectively, to guarantee the stability of the Euler-Maruyama discretisation. Large values of γ and γ' increase the convergence speed at the expense of a higher asymptotic bias (see Theorem 2). However, in most of the experiments we have tested so far, the discretisation bias is highly dominated by the smoothing bias introduced by the Moreau-Yosida regularisation controlled by λ or λ', so our recommendation is to always choose γ and γ' to be as high as possible. A good criterion is to set $\gamma = 0.98/(\mathrm{L}_y + \lambda^{-1})$ and $\gamma' = 0.98\lambda'$.

Selecting $(\delta_n)_{n\in\mathbb{N}^*}$ Regarding the choice of the sequence $(\delta_n)_{n\in\mathbb{N}^*}$, a standard choice is to set for any $n \in \mathbb{N}^*$, $\delta_n = \alpha n^{-\beta}/d$ for some $\alpha > 0$ and $\beta \in [0.6, 0.9]$. In our experiments we use $\beta = 0.8$. For α we recommend, for the case where θ is scalar, to start with $\alpha = 1/\theta_0$ and then adjust if necessary. The scale should be adapted such that after the first few iterations, the updates of θ fall within the desired interval. When θ is not scalar, it is sometimes better to use different scales for every component of θ. In order to speed-up the algorithm it might be useful to perform the estimation on a logarithmic scale, *i.e.* in the scalar case instead of maximizing $\theta \mapsto p(y|\theta)$ we aim at maximizing $\eta \mapsto p(y|\mathrm{e}^{\eta})$ and set $\theta = \mathrm{e}^{\eta}$.

B.2 Testing the MCMC sampler

Before trying to adjust the value of $\theta \in \Theta$ with the algorithm, we strongly recommend to start by testing the MCMC sampler with a fixed value of θ. A simple way to see if the Markov chain is working as expected, is to plot the value of the log-probability of the samples.

As mentioned in Section 2, there is a useful concentration phenomenon studied in [15, Theorem 1.2] which implies that for high-dimensional log-concave densities π, a Markov chain targeting π eventually start generating samples X_n for which $\log \pi(X_n)$ is approximately constant (and close to the entropy). Therefore, if the MCMC sampling is successful the log-probability stabilizes after some iterations and remains more or less constant.



(a) TV-deblurring (b) TGV-denoising

Figure 21: Evolution of $(\log p(X_n|y,\theta))_{n\in\mathbb{N}}$ with $(X_n)_{n\in\mathbb{N}}$ sampled using MYULA and targeting $p(\cdot|y,\theta)$. Results for (a) TV-deblurring with SNR = 40dB and (b) TGV-denoising with SNR = 8dB.

Conversely, if plots show that the chain is divergent or very unstable, then there might be a problem with the sampler. A common cause for divergence is setting a discretisation step-size that is too large. We would advice not to proceed with the estimation of θ until the sampler shows a stable behaviour similar to the ones shown on Figure 21.

B.3 Working with two MCMC chains in Algorithm 3

Using two MCMC kernels simultaneously can be problematic if their convergence speed, or effective sample size per iteration, is very dissimilar as this will degrade the converge properties of the SAPG algorithm. This can be mitigated by adding some "thinning" to the slower MCMC kernel, which essentially means that we concatenate several iterations of the that kernel to artificially improve its performance.

To diagnose imbalances in the Markov kernels we use the fact that the residual $\|g(X_n) - g(\bar{X}_n)\|$ should vanish as n increases, i.e., $g(X_n)$ will become close to $g(\bar{X}_n)$. It is therefore useful to plot the traces of $(g(X_n))_{n\in\mathbb{N}}$ and $(g(\bar{X}_n))_{n\in\mathbb{N}}$ to check that the algorithm is converging. This is illustrated in

Figure 22 below, where we observe how these terms become closer as the number of iterations increases.

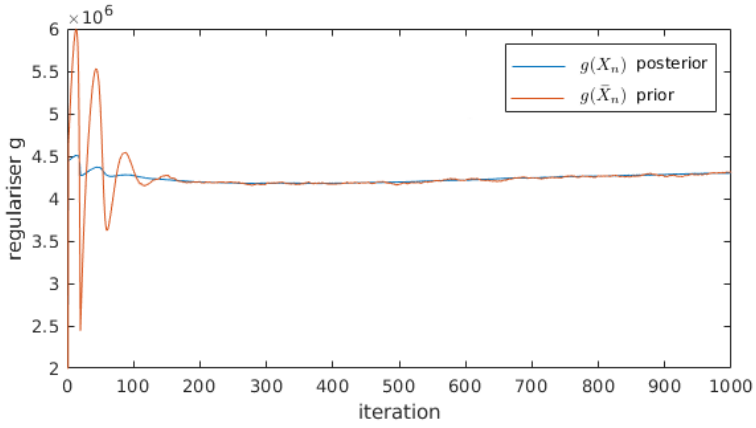


Figure 22: Evolution of the iterates $(g(X_n))_{n\in\mathbb{N}}$ and $(g(\bar{X}_n))_{n\in\mathbb{N}}$ for the proposed method in a deblurring experiment with a TV prior and SNR = 40dB.

If $\|g(X_n) - g(\bar{X}_n)\|$ does not vanish as n increases this could indicate that the step δ_n is not properly chosen, or that there is a significant difference in the speed of the two MCMC kernels. To diagnose the later, we plot the sample autocorrelation for each chain using g as summary statistic. If the autocorrelation plots decay at significantly different speeds then it is necessary to reduce the correlation within the slower chain by either introducing some thinning or by increasing the step-size γ. Usually, we observe that the slowest chain is the one coming from the prior distribution, *i.e.* from the family of kernels $\{\bar{\mathrm{R}}_{\gamma',\theta} : \gamma' \in (0, \bar{\gamma}], \theta \in \Theta\}$. In that case, we increase the step-size γ' to improve convergence speed, at the expense of some additional bias. Note that increasing γ' will typically require to also increase λ'.

B.4 Convergence speed

The bottleneck in convergence speed is the correlation between the samples generated by the MCMC kernels. To increase the convergence speed one has two main alternatives: a) to reduce the correlation between samples, or b) to reduce the computational cost of each iteration in order to afford more iterations.

Reducing sample correlation To reduce the correlation between samples, the step-size γ has to be as large as possible. If running the algorithm with two chains, and the kernel sampling from the prior distribution is the limiting factor, one can consider increasing the smoothing parameter λ' of this particular kernel, in order to be able to increase the value of the discretisation step-size γ'. In more general cases where the limiting factor for γ is L_y there are a few strategies that might help overcome this difficulty. The first strategy is to use preconditioning (see the hyperspectral unmixing experiment in Section 5.3). This is a standard practice which consists of re-scaling the problem to reduce anisotropy and improve the condition number.

Speeding up each iteration The most computationally heavy step in a MYULA iteration is usually the evaluation of the proximal operator. If the proximal operator is being approximated by an iterative solver, it is worth trying to improve efficiency by either using better solver, by warm starting iterations, or by using a weaker convergence criterion.

B.5 Estimation Bias

If the algorithm converges but towards a poor value of $\theta \in \Theta$ it might be due to the bias in the MCMC kernels. As mentioned previously, there are many levels of approximation and the bias is mostly affected by the discretisation step γ and the smoothing parameter λ. However, based on what we have observed in practice, the limiting factor tends to be λ. If there is a bias issue, we recommend trying to reduce λ to obtain a better approximation of the target distribution. This will typically lead to a degradation in convergence speed, as it increases the Lipschitz constant of the gradient and thus lowers the bound for the maximum step-size γ. If the step size becomes too small, then the convergence speed can be dramatically reduced. When convergence is slowed down, special attention has to be paid in the case of the double MCMC chain algorithm. If the effective sample size of the two chains becomes too dissimilar, the algorithm might have difficulty converging. In this case, it is possible to do some thinning in the slower chain, as suggested in Appendix B.3.

C Fair comparison of different methodologies

Comparing different techniques for selecting the value of the regularisation parameter is not as simple as it might seem at first sight. Some algorithms such as SUGAR, are solver dependent and try to find the best value of θ for a given solver, with a given setup (number of iterations, parameters, etc.). Other algorithms, such as the hierarchical one proposed in [64] depend on the solver, but do not seek to optimize θ for that solver but rather for a general case. The algorithm we propose does not depend directly on the solver.

When running statistics on our experiments we noticed an interesting phenomenon. For the deblurring experiments, we use the solver SALSA [1], which is an efficient implementation of the alternating direction method of multipliers (ADMM). When running the hierarchical Bayesian algorithm, we implement it with SALSA and set up the tolerance to 10^{-3} and 150 iterations which seemed sufficient to render very good results. However, when we build the MSE(θ) curves for Figure 10 (by sampling many points and interpolating), we use SALSA with tolerance 10^{-5} and 1000 iterations as there were some pathological values of θ for which SALSA did not converge well with tolerance 10^{-3}. As it may be seen on Figure 23, the position of the minimum MSE changes for the two different SALSA configurations. When computing the average results for 10 images, the parameters obtained with the hierarchical method fell closer to the minimum of the red curve, and the ones obtained with the proposed empirical method fell closer to the minimum of the blue curve. Running the hierarchical method again with tolerance 10^{-5}, the estimated parameters do not change much but the computing times were significantly increased.

The criterion we opt for was to use SALSA with the strictest tolerance and highest number of iterations, because this configuration gives the overall best estimations.

C.1 Comparing with solver-dependent methods

As mentioned previously, algorithms like SUGAR try to find the best value of θ for a given solver, with a given number of iterations, and specific parameters. This means that unless SUGAR is implemented with the exact same solver used to construct the MSE(θ) curves as the ones in Figure 23, the values of θ computed with SUGAR might yield bad results according to the MSE(θ) curve but good results with the specific solver used in SUGAR. For this reason, to achieve a more fair comparison, we compute an equivalent θ_{EQ} in the following way. The SUGAR algorithm returns an estimated θ_{SUG} and a corresponding MSE_{SUG} obtained with that θ_{SUG}. Given an MSE(θ) curve, we define the equivalent θ_{EQ} as

$$\theta_{\text{EQ}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, |\theta - \theta_{\text{SUG}}| \quad s.t. \quad \text{MSE}(\theta) = \text{MS}E_{\text{SUG}} \, .$$

This θ_{EQ} is what we plot in Figure 5. For the lowest SNR value, θ_{EQ} and θ_{SUG} did not differ much in our experiments. However for the other SNR values, the values of θ_{SUG} were significantly smaller than θ_{EQ}.

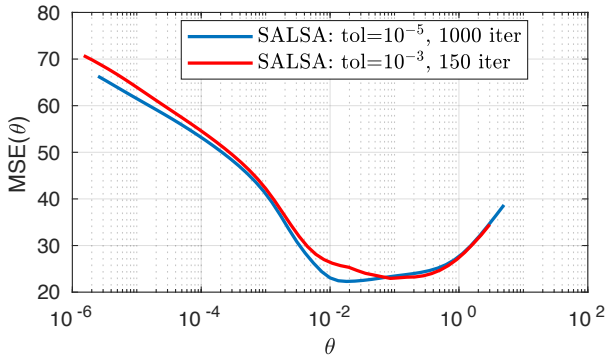


Figure 23: MSE(θ) for wavelet synthesis-ℓ_1 deconvolution for SNR = 20dB with `boat` test image. The curves are computed with different tolerance and maximum iterations using SALSA solver.

D Postponed proofs

In this Section, we show Theorem 1 by applying [80, Theorem 1, Theorem 3], which boils down to verifying that [80, H1, H2] are satisfied. First in Appendix D.2, we show that [80, H1, H2] hold if the sequence is given by (23) where $\{(\mathrm{K}_{\gamma,\theta}, \bar{\mathrm{K}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ defined in (22), *i.e.* we consider PULA as a sampling scheme in the optimization algorithm. Second in Appendix D.3 we check that [80, H1, H2] are satisfied when $\{(\mathrm{K}_{\gamma,\theta}, \bar{\mathrm{K}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(\mathrm{S}_{\gamma,\theta}, \mathrm{S}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ defined in (21), *i.e.* when considering MYULA as a sampling scheme. Finally, we prove the theorem in Appendix D.4.

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

$$\mathrm{R}W(x) \leqslant \lambda W(x) + b \, .$$

D.1 Technical lemmas

We will use the following result.

Lemma 4. *Let* R *be a Markov kernel on* $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ *which satisfies* $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, b\gamma)$ *with* $\lambda \in (0,1)$, $b \geqslant 0$, $\gamma > 0$ *and a measurable function* $W : \mathbb{R}^d \to [1, +\infty)$. *Then, we have for any* $x \in \mathbb{R}^d$

$$\mathrm{R}^{\lceil 1/\gamma \rceil} W(x) \leqslant (1 + b \log^{-1}(1/\lambda) \lambda^{-\bar{\gamma}}) W(x) \, .$$

Proof. Using [80, Lemma 9] we have for any $x \in \mathbb{R}^d$

$$\mathrm{R}^{\lceil 1/\gamma \rceil} W(x) \leqslant \left(\lambda^{\gamma \lceil 1/\gamma \rceil} + b\gamma \sum_{k=0}^{\lceil 1/\gamma \rceil - 1} \lambda^{\gamma k} \right) W(x) \leqslant (1 + b \log^{-1}(1/\lambda) \lambda^{-\bar{\gamma}}) W(x) \, .$$

□

We continue this section by giving some results on proximal operators. Some of them are well but their proof is given for completeness.

Lemma 5. *Let* $\kappa > 0$ *and* $U : \mathbb{R}^d \to \mathbb{R}$ *convex. Assume that* U *is* M*-Lipschitz with* $M \geqslant 0$, *then* U^κ *is* M*-Lipschitz and for any* $x \in \mathbb{R}^d$, $\|x - \mathrm{prox}_U^\kappa(x)\| \leqslant \kappa M$.

Proof. Let $\kappa > 0$. We have for any $x, y \in \mathbb{R}^d$ by (14) and (5)

$$\begin{aligned}
&U^\kappa(x) - U^\kappa(y) \\
&= \|x - \mathrm{prox}_U^\kappa(x)\|^2 / (2\kappa) + U(\mathrm{prox}_U^\kappa(x)) - \|y - \mathrm{prox}_U^\kappa(y)\|^2 / (2\kappa) - U(\mathrm{prox}_U^\kappa(y)) \\
&\leqslant \|y - \mathrm{prox}_U^\kappa(y)\|^2 / (2\kappa) + U(x - y + \mathrm{prox}_U^\kappa(y)) - \|y - \mathrm{prox}_U^\kappa(y)\|^2 / (2\kappa) - U(\mathrm{prox}_U^\kappa(y)) \\
&\leqslant M \|x - y\| \, .
\end{aligned}$$

Hence, U^κ is M-Lipschitz. Since by [11, Proposition 12.30], U^κ is continuously differentiable we have for any $x \in \mathbb{R}^d$, $\|\nabla U^\kappa(x)\| \leqslant M$. Combining this result with the fact that for any $x \in \mathbb{R}^d$, $\nabla U^\kappa(x) = (x - \mathrm{prox}_U^\kappa(x))/\kappa$ by [11, Proposition 12.30] concludes the proof. □

Lemma 6. *Let $U : \mathbb{R}^d \to [0, +\infty)$ be a convex and M-Lipschitz function with $M \geqslant 0$. Then for any $\kappa > 0$ and $z, z' \in \mathbb{R}^d$,*

$$\langle \mathrm{prox}_U^\kappa(z) - z, z \rangle \leqslant -\kappa U(z) + \kappa^2 M^2 + \kappa \{U(z') + M \|z'\|\} \ .$$

Proof. $\kappa > 0$ and $z, z' \in \mathbb{R}^d$. Since $(z - \mathrm{prox}_U^\kappa(z))/\kappa \in \partial U(\mathrm{prox}_U^\kappa(z))$ [11, Proposition 16.44], we have

$$\begin{aligned}
\kappa \{U(z') - U(\mathrm{prox}_U^\kappa(z))\} &\geqslant \langle z - \mathrm{prox}_U^\kappa(z), z' - \mathrm{prox}_U^\kappa(z) \rangle \\
&\geqslant \langle z - \mathrm{prox}_U^\kappa(z), z' - z \rangle + \|z - \mathrm{prox}_U^\kappa(z)\|^2 \\
&\geqslant \langle z - \mathrm{prox}_U^\kappa(z), z' - z \rangle \ .
\end{aligned}$$

Combining this result, the fact that U is M-Lipschitz and Lemma 5 we get that

$$\begin{aligned}
\langle \mathrm{prox}_U^\kappa(z) - z, z \rangle &\leqslant \kappa U(z') - \kappa U(z) + \kappa M \|z - \mathrm{prox}_U^\kappa(z)\| + \|z'\| \|z - \mathrm{prox}_U^\kappa(z)\| \\
&\leqslant -\kappa U(z) + \kappa^2 M^2 + \kappa \{U(z') + M \|z'\|\} \ ,
\end{aligned}$$

which concludes the proof □

Lemma 7. *Let $\kappa_1, \kappa_2 > 0$ and $U : \mathbb{R}^d \to \mathbb{R}$ convex and lower semi-continuous. For any $x \in \mathbb{R}^d$ we have*

$$\|\mathrm{prox}_U^{\kappa_1}(x) - \mathrm{prox}_U^{\kappa_2}(x)\|^2 \leqslant 2(\kappa_1 - \kappa_2)(U(\mathrm{prox}_U^{\kappa_2}(x)) - U(\mathrm{prox}_U^{\kappa_1}(x))) \ .$$

If in addition, U is $\mathtt{M}$-Lipschitz with $M \geqslant 0$ then

$$\|\mathrm{prox}_U^{\kappa_1}(x) - \mathrm{prox}_U^{\kappa_2}(x)\| \leqslant 2M |\kappa_1 - \kappa_2| \ .$$

Proof. By definition of $\mathrm{prox}_U^{\kappa_1}(x)$ we have

$$2\kappa_1 U(\mathrm{prox}_U^{\kappa_1}(x)) + \|x - \mathrm{prox}_U^{\kappa_1}(x)\|^2 \leqslant 2\kappa_1 U(\mathrm{prox}_U^{\kappa_2}(x)) + \|x - \mathrm{prox}_U^{\kappa_2}(x)\|^2 \ .$$

Combining this result and the fact that $(x - \mathrm{prox}_U^{\kappa_2}(x))/\kappa_2 \in \partial U(\mathrm{prox}_U^{\kappa_2}(x))$ we have

$$\begin{aligned}
&\|\mathrm{prox}_U^{\kappa_1}(x) - \mathrm{prox}_U^{\kappa_1}(x)\|^2 \\
&\quad \leqslant 2\kappa_1 \{U(\mathrm{prox}_U^{\kappa_2}(x)) - U(\mathrm{prox}_U^{\kappa_1}(x))\} + 2\langle x - \mathrm{prox}_U^{\kappa_2}(x), \mathrm{prox}_U^{\kappa_1}(x) - \mathrm{prox}_U^{\kappa_2}(x) \rangle \\
&\quad \leqslant 2\kappa_1 \{U(\mathrm{prox}_U^{\kappa_2}(x)) - U(\mathrm{prox}_U^{\kappa_1}(x))\} + 2\kappa_2 \{U(\mathrm{prox}_U^{\kappa_1}(x)) - U(\mathrm{prox}_U^{\kappa_2}(x))\} \\
&\quad \leqslant 2(\kappa_1 - \kappa_2)(U(\mathrm{prox}_U^{\kappa_2}(x)) - U(\mathrm{prox}_U^{\kappa_1}(x))) \ ,
\end{aligned}$$

which concludes the proof. □

Lemma 8. *Let $V : \mathbb{R}^d \to \mathbb{R}$ $\mathtt{m}$-convex and continuously differentiable with $\mathtt{m} \geqslant 0$. Assume that there exits $M \geqslant 0$ such that for any $x, y \in \mathbb{R}^d$*

$$\|\nabla V(x) - \nabla V(y)\| \leqslant M \|x - y\| \ .$$

Assume that there exists $x^\star \in \arg\min_{\mathbb{R}^d} V$, then for any $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < 2/(M + \mathtt{m})$ and $x \in \mathbb{R}^d$

$$\|x - \gamma \nabla V(x)\|^2 \leqslant (1 - \gamma\kappa) \|x\|^2 + \gamma\{(2/(\mathtt{m} + M) - \bar{\gamma})^{-1} + 4\kappa\} \|x^\star\|^2 \ ,$$

with $\kappa = \mathtt{m}M/(\mathtt{m} + M)$.

Proof. Let $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$ and $\bar{\gamma} < 2/(\mathtt{m}+M)$. Using [61, Theorem 2.1.11] and the fact that for any $a, b, \varepsilon > 0$, $\varepsilon a^2 + b^2/\varepsilon \geqslant 2ab$ we have

$$\begin{aligned}
&\|x - \gamma \nabla V(x)\|^2 \\
&\quad \leqslant \|x\|^2 - 2\gamma \langle \nabla V(x) - \nabla V(x^\star), x - x^\star \rangle + \gamma\bar{\gamma} \|\nabla V(x) - \nabla V(x^\star)\|^2 \\
&\qquad + 2\gamma \|x^\star\| \|\nabla V(x) - \nabla V(x^\star)\| \\
&\quad \leqslant \|x\|^2 - 2\gamma\mathtt{\kappa} \|x - x^\star\|^2 - \gamma(2/(\mathtt{m}+M) - \bar{\gamma}) \|\nabla V(x) - \nabla V(x^\star)\|^2 \\
&\qquad + 2\gamma \|x^\star\| \|\nabla V(x) - \nabla V(x^\star)\| \\
&\quad \leqslant \|x\|^2 - 2\gamma\mathtt{\kappa} \|x - x^\star\|^2 - \gamma(2/(\mathtt{m}+M) - \bar{\gamma}) \|\nabla V(x) - \nabla V(x^\star)\|^2 \\
&\qquad + \gamma(2/(\mathtt{m}+M) - \bar{\gamma}) \|\nabla V(x) - \nabla V(x^\star)\|^2 + \gamma/(2/(\mathtt{m}+M) - \bar{\gamma}) \|x^\star\|^2 \\
&\quad \leqslant (1 - 2\gamma\mathtt{\kappa}) \|x\|^2 + 4\gamma\mathtt{\kappa} \|x^\star\| \|x\| + \gamma/(2/(\mathtt{m}+M) - \bar{\gamma}) \|x^\star\|^2 \\
&\quad \leqslant (1 - \gamma\mathtt{\kappa}) \|x\|^2 + \gamma \left\{(2/(\mathtt{m}+M) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} \|x^\star\|^2 \ .
\end{aligned}$$

□

Lemma 9. *Assume* **H1** *and* **H2**. *Then for any $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$ and $x \in \mathbb{R}^d$, we have*

$$\begin{aligned}
&\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))\right\|^2 \\
&\qquad \leqslant (1 - \gamma\mathtt{\kappa}/2) \|x\|^2 + \gamma \left[\bar{\gamma}\kappa^2\mathtt{M}^2 + \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} R_{V,1}^2 \ + 2\kappa^2\mathtt{M}^2\mathtt{\kappa}^{-1}\right] \ ,
\end{aligned}$$

with $\mathtt{\kappa} = \mathtt{mL}/(\mathtt{m}+\mathtt{L})$.

Proof. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Using **H1**, **H2**, Lemma 5, Lemma 8, the Cauchy-Schwarz inequality and that for any $\alpha, \beta \geqslant 0$, $\max_{t \in \mathbb{R}}(-\alpha t^2 + 2\beta t) = \beta^2/\alpha$, we have

$$\begin{aligned}
&\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))\right\|^2 \\
&\quad \leqslant (1 - \gamma\mathtt{\kappa}) \left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)\right\|^2 + \gamma \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} \|x_\theta^\star\|^2 \\
&\quad \leqslant (1 - \gamma\mathtt{\kappa}) \left\|x - \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - x\right\|^2 + \gamma \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} R_{V,1}^2 \\
&\quad \leqslant (1 - \gamma\mathtt{\kappa}) \|x\|^2 + \gamma^2\kappa^2\mathtt{M}^2 + 2\gamma\kappa\mathtt{M} \|x\| + \gamma \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} R_{V,1}^2 \\
&\quad \leqslant (1 - \gamma\mathtt{\kappa}/2) \|x\|^2 + \gamma^2\kappa^2\mathtt{M}^2 + \gamma \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} R_{V,1}^2 + 2\gamma\kappa\mathtt{M} \|x\| - \gamma\mathtt{\kappa} \|x\|^2 / 2 \\
&\quad \leqslant (1 - \gamma\mathtt{\kappa}/2) \|x\|^2 + \gamma\bar{\gamma}\kappa^2\mathtt{M}^2 + \gamma \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\mathtt{\kappa}\right\} R_{V,1}^2 + 2\gamma\kappa^2\mathtt{M}^2\mathtt{\kappa}^{-1} \ .
\end{aligned}$$

□

Lemma 10. *Assume* **H1** *and* **H3**. *Then for any $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < 2/\mathtt{L}$ and $x \in \mathbb{R}^d$, we have*

$$\begin{aligned}
\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))\right\|^2 \leqslant \|x\|^2 + \gamma \big[&3\bar{\gamma}\kappa^2\mathtt{M}^2 + 2\kappa\mathtt{c} + 2\kappa(R_{U,2} + \mathtt{M}R_{U,1}) \\
&+ (2/L - \bar{\gamma})^{-1} R_{V,1}^2 - 2\kappa\eta \|x\|\big] \ .
\end{aligned}$$

Proof. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Using **H1**, **H3**, Lemma 5 and Lemma 6 and Lemma 8 we have

$$\begin{aligned}
&\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))\right\|^2 \leqslant \|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)\|^2 + \gamma/(2/\mathtt{L} - \bar{\gamma}) R_{V,1}^2 \\
&\quad \leqslant \|x\|^2 + \gamma^2\kappa^2\mathtt{M}^2 + 2\langle \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - x, x \rangle + \gamma/(2/\mathtt{L} - \bar{\gamma}) R_{V,1}^2 \\
&\quad \leqslant \|x\|^2 + 3\gamma^2\kappa^2\mathtt{M}^2 - 2\gamma\kappa U(x) + 2\gamma\kappa(U(x_\theta^\sharp) + \mathtt{M}\|x_\theta^\sharp\|) + \gamma/(2/\mathtt{L} - \bar{\gamma}) R_{V,1}^2 \\
&\quad \leqslant \|x\|^2 + 3\gamma^2\kappa^2\mathtt{M}^2 - 2\gamma\kappa\eta \|x\| + 2\gamma\kappa\mathtt{c} \\
&\qquad + 2\gamma\kappa(U(x_\theta^\sharp) + \mathtt{M}\|x_\theta^\sharp\|) + \gamma/(2/\mathtt{L} - \bar{\gamma}) R_{V,1}^2 \\
&\quad \leqslant \|x\|^2 + \gamma \left[3\bar{\gamma}\kappa^2\mathtt{M}^2 + 2\kappa\mathtt{c} + 2\kappa(R_{U,2} + \mathtt{M}R_{U,1}) + (2/L - \bar{\gamma})^{-1} R_{V,1}^2 - 2\kappa\eta \|x\|\right] \ .
\end{aligned}$$

□

Lemma 11. *Assume* H1 *and* H2. *Then for any* $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$ *and* $x \in \mathbb{R}^d$, *we have*

$$\begin{aligned}\left\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\kappa}(x)\right\|^2 &\leqslant (1-\gamma\boldsymbol{\kappa}/2)\left\|x\right\|^2 \\ &\quad + \gamma\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2 + 2\gamma^2\mathtt{ML}R_{V,1} + \gamma^2\mathtt{M}^2 + 2\gamma\mathtt{M}^2(1+\bar{\gamma}\mathtt{L})^2\boldsymbol{\kappa}^{-1}\,,\end{aligned}$$

with $\boldsymbol{\kappa} = \mathtt{mL}/(2\mathtt{m}+2\mathtt{L})$.

Proof. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Using H1, H2, Lemma 5, Lemma 8 and that for any $\alpha, \beta \geqslant 0$, $\max(-\alpha t^2 + 2\beta t) = \beta^2/\alpha$ we have

$$\begin{aligned}&\left\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\kappa}(x)\right\|^2 \\ &\quad\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 + 2\gamma\mathtt{M}\left\|x - \gamma\{\nabla_x V_\theta(x) - \nabla_x V_\theta(x_\theta^\star)\}\right\| + \gamma^2\mathtt{M}^2 \\ &\quad\leqslant (1-\gamma\boldsymbol{\kappa})\left\|x\right\|^2 + \gamma\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\}\left\|x_\theta^\star\right\|^2 \\ &\qquad + 2\gamma\mathtt{M}\left\|x\right\| + 2\gamma^2\mathtt{M}\left\|\nabla_x V_\theta(x) - \nabla_x V_\theta(x_\theta^\star)\right\| + \gamma^2\mathtt{M}^2 \\ &\quad\leqslant (1-\gamma\boldsymbol{\kappa})\left\|x\right\|^2 + \gamma\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\}\left\|x_\theta^\star\right\|^2 \\ &\qquad + 2\gamma\mathtt{M}\left\|x\right\| + 2\gamma^2\mathtt{ML}\left\|x\right\| + 2\gamma^2\mathtt{ML}\left\|x_\theta^\star\right\| + \gamma^2\mathtt{M}^2 \\ &\quad\leqslant (1-\gamma\boldsymbol{\kappa}/2)\left\|x\right\|^2 + \gamma\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2 \\ &\qquad + 2\gamma^2\mathtt{ML}R_{V,1} + \gamma^2\mathtt{M}^2 + 2\gamma\mathtt{M}(1+\bar{\gamma}\mathtt{L})\left\|x\right\| - \gamma\boldsymbol{\kappa}\left\|x\right\|^2/2 \\ &\quad\leqslant (1-\gamma\boldsymbol{\kappa}/2)\left\|x\right\|^2 + \gamma\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2 \\ &\qquad + 2\gamma^2\mathtt{ML}R_{V,1} + \gamma^2\mathtt{M}^2 + 2\gamma\mathtt{M}^2(1+\bar{\gamma}\mathtt{L})^2\boldsymbol{\kappa}^{-1}\,.\end{aligned}$$

□

Lemma 12. *Assume* H1 *and* H3. *Then for any* $\kappa > 0$, $\theta \in \Theta$, $x \in \mathbb{R}^d$ *and* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(2/\mathtt{L}, \eta/(2\mathtt{ML}))$, *we have*

$$\begin{aligned}&\left\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\kappa}(x)\right\|^2 \\ &\quad\leqslant \left\|x\right\|^2 + \gamma\left[(2/L - \bar{\gamma})^{-1}R_{V,1}^2 + 3\bar{\gamma}\mathtt{M}^2 + 2\mathtt{c} + 2(\mathtt{M}R_{U,1} + R_{U,2}) + 2\bar{\gamma}\mathtt{ML}R_{V,2} - \eta\left\|x\right\|\right]\,.\end{aligned}$$

Proof. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Using H1, H3, (14), Lemma 5 and Lemma 6 we have

$$\begin{aligned}&\left\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\kappa}(x)\right\|^2 \\ &\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 - 2\gamma\langle x - \gamma\nabla_x V_\theta(x), \nabla_x U_\theta^{\gamma\kappa}(x)\rangle + \gamma^2\mathtt{M}^2 \\ &\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 - 2\kappa^{-1}\langle x - \gamma\nabla_x V_\theta(x), x - \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)\rangle + \gamma^2\mathtt{M}^2 \\ &\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 - 2\kappa^{-1}\langle x, x - \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)\rangle + 2\kappa^{-1}\gamma\left\|\nabla_x V_\theta(x)\right\|\left\|x - \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)\right\| + \gamma^2\mathtt{M}^2 \\ &\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 + 3\gamma^2\mathtt{M}^2 - 2\gamma\eta\left\|x\right\| + 2\gamma\mathtt{c} + 2\gamma(\mathtt{M}\|x_\theta^\sharp\| + U(x_\theta^\sharp)) + 2\gamma\bar{\gamma}\mathtt{M}\left\|\nabla_x V_\theta(x)\right\| \\ &\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 + 3\gamma\bar{\gamma}\mathtt{M}^2 - 2\gamma\eta\left\|x\right\| \\ &\qquad + 2\gamma\mathtt{c} + 2\gamma(\mathtt{M}R_{U,1} + R_{U,2}) + 2\gamma\bar{\gamma}\mathtt{ML}\left\|x\right\| + 2\gamma\bar{\gamma}\mathtt{ML}\left\|x_\theta^\star\right\| \\ &\leqslant \left\|x - \gamma\nabla_x V_\theta(x)\right\|^2 + 3\gamma\bar{\gamma}\mathtt{M}^2 - \gamma\eta\left\|x\right\| + 2\gamma\mathtt{c} + 2\gamma(\mathtt{M}R_{U,1} + R_{U,2}) + 2\gamma\bar{\gamma}\mathtt{ML}\left\|x_\theta^\star\right\|\,,\end{aligned}$$

where we have used for the last inequality that $\bar{\gamma} < \eta/(2\mathtt{ML})$. Then, we can conclude using H1 and Lemma 8 that

$$\begin{aligned}&\left\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\kappa}(x)\right\|^2 \\ &\leqslant \left\|x\right\|^2 + \gamma/(2/L - \bar{\gamma})R_{V,1}^2 + 3\gamma\bar{\gamma}\mathtt{M}^2 - \gamma\eta\left\|x\right\| + 2\gamma\mathtt{c} + 2\gamma(\mathtt{M}R_{U,1} + R_{U,2}) + 2\gamma\bar{\gamma}\mathtt{ML}R_{V,1} \\ &\leqslant \left\|x\right\|^2 + \gamma\left[(2/L - \bar{\gamma})^{-1}R_{V,1}^2 + 3\bar{\gamma}\mathtt{M}^2 + 2\mathtt{c} + 2(\mathtt{M}R_{U,1} + R_{U,2}) + 2\bar{\gamma}\mathtt{ML}R_{V,2} - \eta\left\|x\right\|\right]\,.\end{aligned}$$

□

D.2 Checking [80, H1, H2] for PULA

We show that under H2 or H3, Foster-Lyapunov drifts hold for PULA in Lemma 13 and Lemma 14. This implies that [80, H1a] holds, see Lemma 15. Combining these Foster-Lyapunov drifts with an appropriate minorization condition Lemma 16, we obtain the geometric ergodicity of the underlying Markov chain in Theorem 17, which implies [80, H1b]. Then, we show that the distance between the invariant probability distribution of the Markov chain and the target distribution is controlled in Corollary 22 and therefore [80, H1c] is satisfied. Finally, we show that [80, H2] is satisfied in Proposition 23.

Lemma 13. *Assume H1 and H2. Then for any $\kappa > 0$, $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$, $\mathrm{S}_{\gamma,\theta}$ and $\bar{\mathrm{S}}_{\gamma,\theta}$ satisfy $\mathbf{D}_{\mathrm{d}}(W_2, \lambda_2^\gamma, b_2\gamma)$ with*

$$\begin{aligned}
&\lambda_2 = \exp\left[-\boldsymbol{\kappa}/2\right] , \\
&b_2 = \bar{\gamma}\kappa^2\mathtt{M}^2 + \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,2}^2 + 2d + 2\kappa^2\mathtt{M}^2\boldsymbol{\kappa}^{-1} + \boldsymbol{\kappa}/2 , \\
&\boldsymbol{\kappa} = \mathtt{m}\mathtt{L}/(\mathtt{m}+\mathtt{L}) .
\end{aligned}$$

In addition, for any $m \in \mathbb{N}^*$, there exist $\lambda_m \in (0,1)$, $b_m \geqslant 0$ such that for any $\kappa > 0$, $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$, $\mathrm{S}_{\gamma,\theta}$ and $\bar{\mathrm{S}}_{\gamma,\theta}$ satisfy $\mathbf{D}_{\mathrm{d}}(W_m, \lambda_m^\gamma, b_m\gamma)$.

Proof. We show the property for $\mathrm{S}_{\gamma,\theta}$ only as the proof for $\bar{\mathrm{S}}_{\gamma,\theta}$ is identical. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Let Z be a d-dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 9 we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \|y\|^2 \mathrm{S}_{\gamma,\theta}(x, \mathrm{d}y) &= \mathbb{E}\left[\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma\nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)) + \sqrt{2\gamma}Z\right\|^2\right] \\
&\leqslant (1 - \gamma\boldsymbol{\kappa}/2)\|x\|^2 + \gamma\left[\bar{\gamma}\kappa^2\mathtt{M}^2 + \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2 \quad + 2\kappa^2\mathtt{M}^2\boldsymbol{\kappa}^{-1}\right] + 2\gamma d .
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\int_{\mathbb{R}^d} (1 + \|y\|^2)\mathrm{S}_{\gamma,\theta}(x, \mathrm{d}y) \leqslant (1 - \gamma\boldsymbol{\kappa}/2)(1 + \|x\|^2) + \gamma\left[\bar{\gamma}\kappa^2\mathtt{M}^2\right. \\
\left. + \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2 + 2d + 2\kappa^2\mathtt{M}^2\boldsymbol{\kappa}^{-1} + \boldsymbol{\kappa}/2\right] ,
\end{aligned}$$

which concludes the first part of the proof using that for any $t \geqslant 0$, $1 - t \leqslant \mathrm{e}^{-t}$.. Let $\mathcal{T}_{\gamma,\theta}(x) = \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma\nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))$. In the sequel, for any $k \in \{1, \ldots, m\}$, $b, \tilde{b}_k \geqslant 0$ and $\lambda, \tilde{\lambda}_k \in [0,1)$ are constants independent of γ which may take different values at each appearance. Note that using Lemma 9, for any $k \in \{1, \ldots, 2m\}$ there exist $\tilde{\lambda}_k \in (0,1)$ and $\tilde{b}_k \geqslant 0$ such that

$$\begin{aligned}
\|\mathcal{T}_{\gamma,\theta}(x)\|^k &\leqslant \{\tilde{\lambda}_k^\gamma \|x\| + \gamma\tilde{b}_k\}^k \\
&\leqslant \tilde{\lambda}_k^{\gamma k}\|x\|^k + \gamma 2^k \max(\tilde{b}_k, 1)^k \max(\bar{\gamma}, 1)^{2k-1}\left\{1 + \|x\|^{k-1}\right\} \\
&\leqslant \tilde{\lambda}_k^\gamma \|x\|^k + \tilde{b}_k\gamma\left\{1 + \|x\|^{k-1}\right\} \leqslant (1 + \|x\|^k)(1 + \tilde{b}_k\gamma) .
\end{aligned} \tag{30}$$

Therefore, combining (30) and the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d}(1+\|y\|^2)\mathrm{S}_{\gamma,\theta}(x,\mathrm{d}y) = 1+\mathbb{E}\left[(\|\mathcal{T}_{\gamma,\theta}(x)\|^2+2\sqrt{2\gamma}\langle\mathcal{T}_{\gamma,\theta}(x),Z\rangle+2\gamma\|Z\|^2)^m\right]\\
&= 1+\sum_{k=0}^{m}\sum_{\ell=0}^{k}\binom{m}{k}\binom{k}{\ell}\|\mathcal{T}_{\gamma,\theta}(x)\|^{2(m-k)}\,2^{(3k-\ell)/2}\gamma^{(k+\ell)/2}\mathbb{E}\left[\langle\mathcal{T}_{\gamma,\theta}(x),Z\rangle^{k-\ell}\|Z\|^{2\ell}\right]\\
&\leqslant 1+\|\mathcal{T}_{\gamma,\theta}(x)\|^{2m}\\
&\quad+2^{3m/2}\sum_{k=1}^{m}\sum_{\ell=0}^{k}\binom{m}{k}\binom{k}{\ell}\|\mathcal{T}_{\gamma,\theta}(x)\|^{2(m-k)}\,\gamma^{(k+\ell)/2}\mathbb{E}\left[\langle\mathcal{T}_{\gamma,\theta}(x),Z\rangle^{k-\ell}\|Z\|^{2\ell}\right]\mathbb{1}_{\{(1,0)\}^c}(k,\ell)\\
&\leqslant 1+\|\mathcal{T}_{\gamma,\theta}(x)\|^{2m}\\
&\quad+\gamma 2^{3m/2}\sum_{k=1}^{m}\sum_{\ell=0}^{k}\binom{m}{k}\binom{k}{\ell}\|\mathcal{T}_{\gamma,\theta}(x)\|^{2m-k-\ell}\,\bar\gamma^{(k+\ell)/2-1}\mathbb{E}\left[\|Z\|^{k+\ell}\right]\mathbb{1}_{\{(1,0)\}^c}(k,\ell)\\
&\leqslant 1+\lambda_{2m}^{\gamma}\|x\|^{2m}+b_{2m}\gamma\left\{1+\|x\|^{2m-1}\right\}\\
&\quad+\gamma 2^{3m/2}2^{2m}\max(\bar\gamma,1)^{2m}\sup_{k\in\{1,\dots,m\}}\left\{(1+\tilde b_k\bar\gamma)\mathbb{E}\left[\|Z\|^k\right]\right\}(1+\|x\|^{2m-1})\\
&\leqslant 1+\lambda^{\gamma}\|x\|^{2m}+\gamma b(1+\|x\|^{2m-1})\\
&\leqslant \lambda^{\gamma/2}(1+\|x\|^{2m})+\gamma b(1+\|x\|^{2m-1})+\lambda^{\gamma}(1+\|x\|^{2m})-\lambda^{\gamma/2}(1+\|x\|^{2m})\,.
\end{aligned}$$

Using that $\lambda^{\gamma}-\lambda^{\gamma/2}\leqslant-\log(1/\lambda)\gamma\lambda^{\gamma/2}/2$, concludes the proof. □

Lemma 14. *Assume* H1 *and* H3. *Then for any* $\kappa>0$, $\theta\in\Theta$ *and* $\gamma\in(0,\bar\gamma]$ *with* $\bar\gamma<2/\mathtt{L}$, $\mathrm{S}_{\gamma,\theta}$ *and* $\bar{\mathrm{S}}_{\gamma,\theta}$ *satisfy* $\mathbf{D}_{\mathrm{d}}(W,\lambda^{\gamma},b\gamma)$ *with*

$$\begin{aligned}
&\lambda=\mathrm{e}^{-\alpha^2}\,,\\
&b_e=(3/2)\bar\gamma\kappa^2\mathtt{M}^2+\kappa\mathtt{c}+\kappa(R_{U,2}+\mathtt{M}R_{U,1})+(4/L-2\bar\gamma)^{-1}R_{V,1}^2+d+2\alpha\\
&b=\alpha b_e\mathrm{e}^{\alpha\bar\gamma b_e}W(R)\,,\\
&W=W_{\alpha}\,,\qquad 0<\alpha<\kappa\eta/4\,,\\
&R_{\eta}=\max\left(b_e/(\kappa\eta-4\alpha),1\right)\,.
\end{aligned}$$

Proof. We show the property for $\mathrm{S}_{\gamma,\theta}$ only as the proof for $\bar{\mathrm{S}}_{\gamma,\theta}$ is identical. Let $\kappa>0$, $\theta\in\Theta$, $\gamma\in(0,\bar\gamma]$, $x\in\mathbb{R}^d$, and Z be a d-dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 10 we have

$$\begin{aligned}
&\int_{\mathbb{R}^d}\|y\|^2\,\mathrm{S}_{\gamma,\theta}(x,\mathrm{d}y)\leqslant\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)-\gamma\nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))\right\|^2+2\gamma d\\
&\qquad\leqslant\|x\|^2+\gamma\left[3\bar\gamma\kappa^2\mathtt{M}^2+2\kappa\mathtt{c}+2\kappa(R_{U,2}+\mathtt{M}R_{U,1})+(2/L-\bar\gamma)^{-1}R_{V,1}^2+2d-2\kappa\eta\|x\|\right]\,.
\end{aligned}$$

Using the log-Sobolev inequality [9, Proposition 5.4.1] and Jensen's inequality we get that

$$\begin{aligned}
\mathrm{S}_{\gamma,\theta}W(x)&\leqslant\exp\left[\alpha\,\mathrm{S}_{\gamma,\theta}\phi(x)+\alpha^2\gamma\right]\\
&\leqslant\exp\left[\alpha\left(1+\int_{\mathbb{R}^d}\|y\|^2\,\mathrm{S}_{\gamma,\theta}(x,\mathrm{d}y)\right)^{1/2}+\alpha^2\gamma\right]\,.
\end{aligned}\tag{31}$$

We now distinguish two cases.

(a) If $\|x\|\geqslant R_{\eta}$ then $\phi^{-1}(x)\|x\|\geqslant 1/2$ and $3\bar\gamma\kappa^2\mathtt{M}^2+2\kappa\mathtt{c}+2\kappa(R_{U,2}+\mathtt{M}R_{U,1})+(2/L-\bar\gamma)^{-1}R_{V,1}^2+2d-2\kappa\eta\|x\|\leqslant-8\alpha\|x\|$. In this case using that for any $t\geqslant 0$, $\sqrt{1+t}-1\leqslant t/2$ we get

$$\begin{aligned}
&\left(1+\int_{\mathbb{R}^d}\|y\|^2\,\mathrm{S}_{\gamma,\theta}(x,\mathrm{d}y)\right)^{1/2}-\phi(x)\\
&\qquad\leqslant\gamma\phi^{-1}(x)\left[3\bar\gamma\kappa^2\mathtt{M}^2+2\kappa\mathtt{c}+2\kappa(R_{U,2}+\mathtt{M}R_{U,1})+(2/L-\bar\gamma)^{-1}R_{V,1}^2+2d-2\kappa\eta\|x\|\right]/2\\
&\qquad\leqslant-4\alpha\gamma\phi^{-1}(x)\|x\|\leqslant-2\alpha\gamma\,.
\end{aligned}$$

Hence,

$$\mathrm{S}_{\gamma,\theta}W(x) \leqslant \exp\left[\alpha\left(1+\int_{\mathbb{R}^d}\|y\|^2\,\mathrm{S}_{\gamma,\theta}(x,\mathrm{d}y)\right)^{1/2}+\alpha^2\gamma\right] \leqslant \mathrm{e}^{-\alpha^2\gamma}W(x)\;.$$

(b) If $\|x\| \leqslant R_\eta$ then using that for any $t \geqslant 0$, $\sqrt{1+t}-1 \leqslant t/2$

$$\left(1+\int_{\mathbb{R}^d}\|y\|^2\,\mathrm{S}_{\gamma,\theta}(x,\mathrm{d}y)\right)^{1/2}-\phi(x) \leqslant \gamma\left[(3/2)\bar{\gamma}\kappa^2\mathtt{M}^2+\kappa\mathtt{c}+\kappa(R_{U,2}+\mathtt{M}R_{U,1})+(4/L-2\bar{\gamma})^{-1}R_{V,1}^2+d\right]\;.$$

Therefore we get using (31)

$$\mathrm{S}_{\gamma,\theta}W(x)/W(x) \leqslant \exp\left[\alpha\gamma\left\{(3/2)\bar{\gamma}\kappa^2\mathtt{M}^2+\kappa\mathtt{c}+\kappa(R_{U,2}+\mathtt{M}R_{U,1})+(4/L-2\bar{\gamma})^{-1}R_{V,1}^2+d+\alpha\right\}\right] \leqslant \mathrm{e}^{\alpha b_e\gamma}\;.$$

Since for all $a \geqslant b$, $\mathrm{e}^a-\mathrm{e}^b \leqslant (a-b)\mathrm{e}^a$ we obtain that

$$\mathrm{S}_{\gamma,\theta}W(x) \leqslant \lambda^\gamma W(x)+\gamma\alpha b_e\mathrm{e}^{\alpha\bar{\gamma}b_e}W(R_\eta)\;,$$

which concludes the proof.

□

Lemma 15. *Assume* H1, H2 *or* H3, *and let* $(X_k^n,\bar{X}_k^n)_{n\in\mathbb{N},k\in\{0,\ldots,m_n\}}$ *be given by* (23) *with* $\{(\mathrm{K}_{\gamma,\theta},\bar{\mathrm{K}}_{\gamma,\theta}) : \gamma\in(0,\bar{\gamma}]\,,\theta\in\Theta\} = \{(\mathrm{S}_{\gamma,\theta},\mathrm{S}_{\gamma,\theta}) : \gamma\in(0,\bar{\gamma}]\,,\theta\in\Theta\}$ *and* $\kappa>0$. *Then there exists* $A_1 \geqslant 1$ *such that for any* $n,p\in\mathbb{N}$ *and* $k\in\{0,\ldots,m_n\}$

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{S}_{\gamma_n,\theta_n}^p W(X_k^n)\,|X_0^0\right] \leqslant A_1 W(X_0^0)\;,\\
&\mathbb{E}\left[\bar{\mathrm{S}}_{\gamma_n,\theta_n}^p W(\bar{X}_k^n)\,|\bar{X}_0^0\right] \leqslant A_1 W(\bar{X}_0^0)\;,\\
&\mathbb{E}\left[W(X_0^0)\right]<+\infty\;,\qquad \mathbb{E}\left[W(\bar{X}_0^0)\right]<+\infty\;,
\end{aligned}$$

with $W=W_m$ *with* $m\in\mathbb{N}^*$ *and* $\bar{\gamma}<2/(\mathtt{m}+\mathtt{L})$ *if* H2 *holds and* $W=W_\alpha$ *with* $\alpha<\kappa\eta/4$ *and* $\bar{\gamma}<2/\mathtt{L}$ *if* H3 *holds.*

Proof. Combining [80, Lemma S15] and Lemma 13 if H2 holds or Lemma 14 if H3 holds conclude the proof. □

Lemma 16. *Assume* H1. *For any* $\kappa>0$, $\theta\in\Theta$, $\gamma\in(0,\bar{\gamma}]$ *with* $\bar{\gamma}<2/\mathtt{L}$, *and* $x,y\in\mathbb{R}^d$

$$\max\left(\|\delta_x\mathrm{S}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}-\delta_y\mathrm{S}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\|_{\mathrm{TV}},\|\delta_x\bar{\mathrm{S}}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}-\delta_y\bar{\mathrm{S}}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\|_{\mathrm{TV}}\right) \leqslant 1-2\boldsymbol{\Phi}\left\{-\|x-y\|/(2\sqrt{2})\right\}\;,$$

where $\boldsymbol{\Phi}$ *is the cumulative distribution function of the standard normal distribution on* $\mathbb{R}$.

Proof. We only show that for any $\kappa>0$, $\theta\in\Theta$, $\gamma\in(0,\bar{\gamma}]$ with $\bar{\gamma}<2/\mathtt{L}$, and $x,y\in\mathbb{R}^d$, $\|\delta_x\mathrm{S}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}-\delta_y\mathrm{S}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\|_{\mathrm{TV}} \leqslant 1-2\boldsymbol{\Phi}\left\{-\|x-y\|/(2\sqrt{2})\right\}$ since the proof for $\bar{\mathrm{S}}_{\gamma,\theta}$ is similar. Let $\kappa>0$, $\theta\in\Theta$, $\gamma\in(0,\bar{\gamma}]$ with $\bar{\gamma}<2/\mathtt{L}$. Using [61, Theorem 2.1.5, Equation (2.1.8)] and that the proximal operator is non-expansive [11, Proposition 12.28], we have for any $x,y\in\mathbb{R}^d$

$$\begin{aligned}
&\left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)-\mathrm{prox}_{U_\theta}^{\gamma\kappa}(y)-\gamma(\nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))-\nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(y)))\right\|\\
&\leqslant \left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)-\mathrm{prox}_{U_\theta}^{\gamma\kappa}(y)\right\| \leqslant \|x-y\|\;.
\end{aligned}$$

The proof is then an application of [81, Proposition 3b] with $\ell\leftarrow 1$, for any $x\in\mathbb{R}^d$, $\mathcal{T}_{\gamma,\theta}(x)\leftarrow \mathrm{prox}_{U_\theta}^{\gamma\kappa}(x)-\gamma\nabla_x V_\theta(\mathrm{prox}_{U_\theta}^{\gamma\kappa}(x))$ and $\Pi\leftarrow\mathrm{Id}$. □

Theorem 17. *Assume* H1 *and* H2 *or* H3*. Let* $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$ *if* H2 *holds and* $\bar{\gamma} < 2/\mathtt{L}$ *if* H3 *holds. Then for any* $a \in (0,1]$*, there exist* $A_{2,a} \geqslant 0$ *and* $\rho_a \in (0,1)$ *such that for any* $\kappa > 0$*,* $\theta \in \Theta$*,* $\gamma \in (0, \bar{\gamma}]$*,* $\mathrm{S}_{\gamma,\theta}$ *and* $\bar{\mathrm{S}}_{\gamma,\theta}$ *admit a invariant probability measure* $\pi_{\gamma,\theta}$ *and* $\bar{\pi}_{\gamma,\theta}$ *respectively. In addition, for any* $x, y \in \mathbb{R}^d$ *and* $n \in \mathbb{N}$ *we have*

$$\max\left(\|\delta_x \mathrm{S}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_{W^a}, \|\delta_x \bar{\mathrm{S}}_{\gamma,\theta}^n - \bar{\pi}_{\gamma,\theta}\|_{W^a}\right) \leqslant A_{2,a}\rho_a^{\gamma n} W^a(x) \;,$$
$$\max\left(\|\delta_x \mathrm{S}_{\gamma,\theta}^n - \delta_y \mathrm{S}_{\gamma,\theta}^n\|_{W^a}, \|\delta_x \bar{\mathrm{S}}_{\gamma,\theta}^n - \delta_y \bar{\mathrm{S}}_{\gamma,\theta}^n\|_{W^a}\right) \leqslant A_{2,a}\rho_a^{\gamma n}\left\{W^a(x) + W^a(y)\right\} \;,$$

with $W = W_m$ *and* $m \in \mathbb{N}^*$ *if* H2 *holds and* $W = W_\alpha$ *with* $\alpha < \kappa\eta/4$ *if* H3 *holds.*

Proof. We only show that for any $a \in (0,1]$, there exist $A_{2,a} \geqslant 0$ and $\rho_a \in (0,1)$ such that for any $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ we have $\|\delta_x \mathrm{S}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_{W^a} \leqslant A_{2,a}\rho_a^{\gamma n} W^a(x)$ and $\|\delta_x \mathrm{S}_{\gamma,\theta}^n - \delta_y \mathrm{S}_{\gamma,\theta}^n\|_{W^a} \leqslant A_{2,a}\rho_a^{\gamma n}\left\{W^a(x) + W^a(y)\right\}$, since the proof for $\bar{\mathrm{S}}_{\gamma,\theta}$ is similar . Let $a \in [0,1]$. First, using Jensen's inequality and Lemma 13 if H2 holds or Lemma 14 if H3 holds, we get that there exist λ_a and b_a such that for any $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$, $\mathrm{S}_{\gamma,\theta}$ and $\bar{\mathrm{S}}_{\gamma,\theta}$ satisfy $\mathbf{D}_{\mathrm{d}}(W^a, \lambda_a^\gamma, b_a\gamma)$. Combining [81, Theorem 6], Lemma 16 and $\mathbf{D}_{\mathrm{d}}(W^a, \lambda_a^\gamma, b_a\gamma)$, we get that there exist $\bar{A}_{2,a} \geqslant 0$ and $\rho_a \in (0,1)$ such that for any $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$, $\mathrm{S}_{\gamma,\theta}$ and $\bar{\mathrm{S}}_{\gamma,\theta}$ admit invariant probability measures $\pi_{\gamma,\theta}$ and $\bar{\pi}_{\gamma,\theta}$ respectively and

$$\max\left\{\|\delta_x \mathrm{S}_{\gamma,\theta}^n - \delta_y \mathrm{S}_{\gamma,\theta}^n\|_{W^a}, \|\delta_x \bar{\mathrm{S}}_{\gamma,\theta}^n - \delta_y \bar{\mathrm{S}}_{\gamma,\theta}^n\|_{W^a}\right\} \leqslant \bar{A}_{2,a}\rho_a^{\gamma n}\left\{W^a(x) + W^a(y)\right\} \;. \tag{32}$$

Using that for any $\kappa > 0$, $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, $\mathrm{S}_{\gamma,\theta}$ and $\bar{\mathrm{S}}_{\gamma,\theta}$ satisfy $\mathbf{D}_{\mathrm{d}}(W^a, \lambda_a^\gamma, b_a\gamma)$ and [80, Lemma S2] we have

$$\pi_{\gamma,\theta}(W^a) \leqslant b_a\gamma/(1-\lambda_a^\gamma) \leqslant b_a\lambda_a^{-\bar{\gamma}}/\log(1/\lambda_a) \;. \tag{33}$$

Hence, combining (32) and (33), we have for any $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}$

$$\max\left\{\|\delta_x \mathrm{S}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_W, \|\delta_x \bar{\mathrm{S}}_{\gamma,\theta}^n - \bar{\pi}_{\gamma,\theta}\|_W\right\} \leqslant \bar{A}_{2,a}\rho_a^{\gamma n}(1 + b_a\lambda_a^{-\bar{\gamma}}/\log(1/\lambda_a))W^a(x) \;.$$

We conclude upon letting $A_{2,a} = \bar{A}_{2,a}(1 + b_a\lambda_a^{-\bar{\gamma}}/\log(1/\lambda_a))$. □

Lemma 18. *Assume* H1 *and* H2 *or* H3*. We have* $\sup_{\theta\in\Theta}\{\pi_\theta(W) + \bar{\pi}_\theta(W)\} < +\infty$*, with* $W = W_m$ *with* $m \in \mathbb{N}^*$ *if* H2 *holds and* $W = W_\alpha$ *with* $\alpha < \eta$ *if* H3 *holds.*

Proof. We only show that $\sup_\theta \pi_\theta(W) < +\infty$ since the proof for $\bar{\pi}_\theta$ is similar. Let $m \in \mathbb{N}^*$, $\alpha < \eta$ and $\theta \in \Theta$ The proof is divided into two parts.

(a) If H2 holds then using H1-(b) we have

$$\begin{aligned}
\int_{\mathbb{R}^d} (1 + \|x\|^{2m}) \exp\left[-U_\theta(x) - V_\theta(x)\right] \mathrm{d}x &\leqslant \int_{\mathbb{R}^d} (1 + \|x\|^{2m}) \exp\left[-V_\theta(x)\right] \mathrm{d}x \\
&\leqslant \int_{\mathbb{R}^d} (1 + \|x\|^{2m}) \exp\left[-V_\theta(x_\theta^\star) - \mathtt{m}\left\|x - x_\theta^\star\right\|^2/2\right] \mathrm{d}x \\
&\leqslant \exp\left[R_{V,3} + \mathtt{m}R_{V,1}^2/2\right] \int_{\mathbb{R}^d} (1 + \|x\|^{2m}) \exp\left[\mathtt{m}R_{V,1}\|x\| - \mathtt{m}\|x\|^2/2\right] \mathrm{d}x \;.
\end{aligned}$$

Hence using H1-(a) we have

$$\sup_{\theta\in\Theta} \pi_\theta(W) \leqslant \exp\left[R_{V,3} + \mathtt{m}R_{V,1}^2/2\right] \int_{\mathbb{R}^d} (1 + \|x\|^{2m}) \exp\left[\mathtt{m}R_{V,1}\|x\| - \mathtt{m}\|x\|^2/2\right] \mathrm{d}x \Big/ \inf_{\theta\in\Theta}\left\{\int_{\mathbb{R}^d} \exp\left[-U_\theta(x) - V_\theta(x)\right] \mathrm{d}x\right\} < +\infty \;.$$

(b) if H3 holds then we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \exp\left[\alpha\phi(x)\right] \exp\left[-U_\theta(x) - V_\theta(x)\right] \mathrm{d}x &\leqslant \int_{\mathbb{R}^d} \exp\left[\alpha\phi(x)\right] \exp\left[-U_\theta(x)\right] \mathrm{d}x \\
&\leqslant \mathrm{e}^{\mathtt{c}} \int_{\mathbb{R}^d} \exp\left[\alpha(1 + \|x\|)\right] \exp\left[-\eta\|x\|\right] \mathrm{d}x \;.
\end{aligned}$$

Since $\alpha < \eta$ we have using H1-(a)

$$\sup_{\theta\in\Theta} \pi_\theta(W) \leqslant \mathrm{e}^{\mathtt{c}} \int_{\mathbb{R}^d} \exp\left[\alpha(1+\|x\|)\right] \exp\left[-\eta\,\|x\|\right] \mathrm{d}x \Big/ \inf_{\theta\in\Theta} \left\{ \int_{\mathbb{R}^d} \exp\left[-U_\theta(x) - V_\theta(x)\right] \mathrm{d}x \right\} < +\infty \ ,$$

which concludes the proof. □

Theorem 19. *Assume* H1 *and* H2 *or* H3*. Let* $\bar\gamma < 2/(\mathtt{m}+\mathtt{L})$ *if* H2 *holds and* $\bar\gamma < 2/\mathtt{L}$ *if* H3 *holds. Then for any* $\kappa > 0$*,* $\theta \in \Theta$ *and* $\gamma \in (0, \bar\gamma]$ *we have*

$$\max\left(\|\pi^\sharp_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}}, \|\bar\pi^\sharp_{\gamma,\theta} - \bar\pi_\theta\|_{W^{1/2}} \right) \leqslant \tilde{\mathbf{\Psi}}(\gamma) \ ,$$

where for any $\theta \in \Theta$ *and* $\gamma \in (0, \bar\gamma]$*,* $\pi^\sharp_{\gamma,\theta}$*, respectively* $\bar\pi^\sharp_{\gamma,\theta}$*, is the invariant probability measure of* $\mathrm{S}_{\gamma,\theta}$*, respectively* $\bar{\mathrm{S}}_{\gamma,\theta}$*, given by* (22) *and associated with* $\kappa = 1$*. In addition, for any* $\gamma \in (0, \bar\gamma]$

$$\tilde{\mathbf{\Psi}}(\gamma) = \sqrt{2}\{b\lambda^{-\bar\gamma}/\log(1/\lambda) + \sup_{\theta\in\Theta} \pi_\theta(W) + \sup_{\theta\in\Theta} \bar\pi_\theta(W)\}^{1/2} (\mathtt{L}d + \mathtt{M}^2)^{1/2} \sqrt{\gamma} \ ,$$

and where $W = W_m$ *with* $m \in \mathbb{N}^*$ *and* λ, b *are given in Lemma 13 if* H2 *holds and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta)$ *and* λ, b *are given in Lemma 14 if* H3 *holds.*

Proof. We only show that for any $\kappa > 0$, $\theta \in \Theta$ and $\gamma \in (0, \bar\gamma]$, $\|\pi^\sharp_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} \leqslant \tilde{\mathbf{\Psi}}(\gamma)$, since the proof of $\|\bar\pi^\sharp_{\gamma,\theta} - \bar\pi_\theta\|_{W^{1/2}} \leqslant \tilde{\mathbf{\Psi}}(\gamma)$ is similar. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar\gamma]$ and $x \in \mathbb{R}^d$ Using Theorem 17 we obtain that $(\delta_x \mathrm{S}^n_{\gamma,\theta})_{n\in\mathbb{N}}$, with $\kappa = 1$, is weakly convergent towards $\pi^\sharp_{\gamma,\theta}$. Using that $\mu \mapsto \mathrm{KL}\left(\mu|\pi_\theta\right)$ is lower semi-continuous for any $\theta \in \Theta$, see [30, Lemma 1.4.3b], and [31, Corollary 18] we get that

$$\mathrm{KL}\left(\pi^\sharp_{\gamma,\theta}|\pi_\theta\right) \leqslant \liminf_{n\to+\infty} \mathrm{KL}\left(n^{-1} \sum_{k=1}^n \delta_x \mathrm{S}^k_{\gamma,\theta} \middle| \pi_\theta \right) \leqslant \gamma(\mathtt{L}d + \mathtt{M}^2) \ .$$

Using a generalized Pinsker inequality, see [32, Lemma 24], Lemma 18 and Lemma 13 if H2 holds or Lemma 14 if H3 holds, we get that

$$\begin{aligned} \|\pi^\sharp_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} &\leqslant \sqrt{2}(\pi^\sharp_{\gamma,\theta}(W) + \pi_\theta(W))^{1/2} \mathrm{KL}\left(\pi^\sharp_{\gamma,\theta}|\pi_\theta\right)^{1/2} \\ &\leqslant \sqrt{2}\{b\lambda^{-\bar\gamma}/\log(1/\lambda) + \sup_{\theta\in\Theta} \pi_\theta(W)\}^{1/2} (\mathtt{L}d + \mathtt{M}^2)^{1/2} \gamma^{1/2} \ , \end{aligned}$$

which concludes the proof. □

Lemma 20. *Assume* H1 *and* H2 *or* H3*. Let* $\bar\gamma < 2/(\mathtt{m}+\mathtt{L})$ *if* H2 *holds and* $\bar\gamma < 2/\mathtt{L}$ *if* H3 *holds. Then there exists* $\bar B_3 \geqslant 0$ *such that for any* $\theta \in \Theta$*,* $\gamma \in (0, \bar\gamma]$*,* $x \in \mathbb{R}^d$ *and* $\kappa_i > 0$ *with* $i \in \{1, 2\}$ *we have*

$$\max\left(\|\delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{1,\gamma,\theta} - \delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{2,\gamma,\theta}\|_{W^{1/2}}, \|\delta_x \bar{\mathrm{S}}^{\lceil 1/\gamma\rceil}_{1,\gamma,\theta} - \delta_x \bar{\mathrm{S}}^{\lceil 1/\gamma\rceil}_{2,\gamma,\theta}\|_{W^{1/2}} \right) \leqslant \bar B_3 \gamma\, |\kappa_1 - \kappa_2|\, W^{1/2}(x) \ .$$

where for any $i \in \{1, 2\}$*,* $\theta \in \Theta$ *and* $\gamma \in (0, \bar\gamma]$*,* $\mathrm{S}_{i,\gamma,\theta}$ *is given by* (22) *and associated with* $\kappa \leftarrow \kappa_i$*, and* $W = W_m$ *with* $m \in \mathbb{N}^*$ *if* H2 *holds. In addition,* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta)$ *if* H3 *holds.*

Proof. We only show that for any $\theta \in \Theta$, $\gamma \in (0, \bar\gamma]$, $x \in \mathbb{R}^d$ and $\kappa_i > 0$ with $i \in \{1, 2\}$ we have $\|\delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{1,\gamma,\theta} - \delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{2,\gamma,\theta}\|_{W^{1/2}} \leqslant \bar B_3 \gamma \left\{\kappa_1 + \kappa_2\right\} W^{1/2}(x)$ since the proof for $\bar{\mathrm{S}}_{1,\gamma,\theta}$ and $\bar{\mathrm{S}}_{2,\gamma,\theta}$ is similar. Let $\theta \in \Theta$, $\gamma \in (0, \bar\gamma]$, $x \in \mathbb{R}^d$ and $\kappa_i > 0$ with $i \in \{1, 2\}$. Using a generalized Pinsker inequality, see [32, Lemma 24], we have

$$\begin{aligned} &\|\delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{1,\gamma,\theta} - \delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{2,\gamma,\theta}\|_{W^{1/2}} \\ &\qquad \leqslant \sqrt{2}(\mathrm{S}^{\lceil 1/\gamma\rceil}_{1,\gamma,\theta} W(x) + \mathrm{S}^{\lceil 1/\gamma\rceil}_{2,\gamma,\theta} W(x))^{1/2} \mathrm{KL}\left(\delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{1,\gamma,\theta} | \delta_x \mathrm{S}^{\lceil 1/\gamma\rceil}_{2,\gamma,\theta}\right)^{1/2} \ . \end{aligned} \tag{34}$$

Using [50, Lemma 4.1] we get that $\mathrm{KL}\left(\delta_x \mathrm{S}_{1,\gamma,\theta}^{\lceil 1/\gamma \rceil} | \delta_x \mathrm{S}_{2,\gamma,\theta}^{\lceil 1/\gamma \rceil}\right) \leqslant \mathrm{KL}\left(\tilde{\mu}_1 | \tilde{\mu}_2\right)$ where setting $T = \gamma \lceil 1/\gamma \rceil$, $\tilde{\mu}_i$, $i \in \{1,2\}$, is the probability measure over $\mathcal{B}(\mathrm{C}([0,T],\mathbb{R}^d))$ which is defined for any $\mathsf{A} \in \mathcal{B}(\mathrm{C}([0,T],\mathbb{R}^d))$ by $\tilde{\mu}_i(A) = \mathbb{P}((X_t^i)_{t\in[0,T]} \in \mathsf{A})$, $i \in \{1,2\}$ and for any $t \in [0,T]$

$$\mathrm{d}X_t^i = b_i(t, (X_s^i)_{s\in[0,T]})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \qquad X_0^i = x \ ,$$

with for any $(\omega_s)_{s\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$ and $t \in [0,T]$

$$b_i(t, (\omega_s)_{s\in[0,T]}) = \sum_{p\in\mathbb{N}} \mathbb{1}_{[p\gamma,(p+1)\gamma)}(t)\mathcal{T}(\mathrm{prox}_{U_\theta}^{\gamma\kappa_i}(\omega_{p\gamma})) \ ,$$

where for any $y \in \mathbb{R}^d$, $\mathcal{T}_{\gamma,\theta}(y) = y - \gamma\nabla_x V_\theta(y)$. Since $(X_t^i)_{t\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$, b_i and b are continuous for any $i \in \{1,2\}$, [54, Theorem 7.19] applies and we obtain that $\tilde{\mu}_1 \ll \tilde{\mu}_2$ and

$$\frac{\mathrm{d}\tilde{\mu}_1}{\mathrm{d}\tilde{\mu}_2}((X_t^1)_{t\in[0,T]}) = \exp\left\{(1/4)\int_0^T \left\|b_1(t,(X_s^1)_{s\in[0,T]}) - b_2(t,(X_s^1)_{s\in[0,T]})\right\|^2 \mathrm{d}t \right.$$
$$\left. +(1/2)\int_0^T \langle b_1(t,(X_s^1)_{s\in[0,T]}) - b_2(t,(X_s^1)_{s\in[0,T]}), \mathrm{d}X_t^1\rangle \right\} \ ,$$

where the equality holds almost surely. As a consequence we obtain that

$$\mathrm{KL}\left(\tilde{\mu}_1|\tilde{\mu}_2\right) = (1/4)\mathbb{E}\left[\int_0^T \left\|b_1(t,(X_s^1)_{s\in[0,T]}) - b_2(t,(X_s^1)_{s\in[0,T]})\right\|^2 \mathrm{d}s\right] \ . \tag{35}$$

In addition, using Lemma 7 we have for any $(\omega_s)_{s\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$ and $t \in [0,T]$

$$\left\|b_1(t,(\omega_s)_{s\in[0,T]}) - b_2(t,(\omega_s)_{s\in[0,T]})\right\|^2 = \left\|\mathcal{T}_{\gamma,\theta}(\mathrm{prox}_{U_\theta}^{\gamma\kappa_1}(\omega_{\gamma\lfloor t/\gamma\rfloor})) - \mathcal{T}_{\gamma,\theta}(\mathrm{prox}_{U_\theta}^{\gamma\kappa_2}(\omega_{\gamma\lfloor t/\gamma\rfloor}))\right\|^2$$
$$\leqslant \left\|\mathrm{prox}_{U_\theta}^{\gamma\kappa_1}(\omega_{\gamma\lfloor t/\gamma\rfloor}) - \mathrm{prox}_{U_\theta}^{\gamma\kappa_2}(\omega_{\gamma\lfloor t/\gamma\rfloor})\right\|^2 \leqslant 4\gamma^2(\kappa_1-\kappa_2)^2\mathtt{M}^2 \ . \tag{36}$$

Combining this result and (35) we get that

$$\mathrm{KL}\left(\delta_x \mathrm{S}_{1,\gamma,\theta}^{\lceil 1/\gamma \rceil} | \delta_x \mathrm{S}_{2,\gamma,\theta}^{\lceil 1/\gamma \rceil}\right) \leqslant (1+\bar{\gamma})\mathtt{M}^2\gamma^2\left|\kappa_1-\kappa_2\right|^2 \ . \tag{37}$$

Combining (37) and (34) we get that

$$\|\delta_x \mathrm{S}_{1,\gamma,\theta}^{\lceil 1/\gamma \rceil} - \delta_x \mathrm{S}_{2,\gamma,\theta}^{\lceil 1/\gamma \rceil}\|_{W^{1/2}}$$
$$\leqslant 2^{1/2}(1+\bar{\gamma})^{1/2}\mathtt{M}(\mathrm{S}_{1,\gamma,\theta}^{\lceil 1/\gamma \rceil}W(x) + \mathrm{S}_{2,\gamma,\theta}^{\lceil 1/\gamma \rceil}W(x))^{1/2}\gamma\left|\kappa_1-\kappa_2\right| \ .$$

We conclude the proof upon using Lemma 4 and Lemma 13 if **H2** holds, or Lemma 14 if **H3** holds. □

Proposition 21. *Assume* **H1** *and* **H2** *or* **H3**. *Let* $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$ *if* **H2** *holds and* $\bar{\gamma} < 2/\mathtt{L}$ *if* **H3** *holds. Then there exists* $B_3 \geqslant 0$ *such that for any* $\theta \in \Theta$, $\gamma \in (0,\bar{\gamma}]$ *and* $\kappa_i > 0$ *with* $i \in \{1,2\}$ *we have*

$$\max\left(\|\pi_{\gamma,\theta}^1 - \pi_{\gamma,\theta}^2\|_{W^{1/2}}, \|\bar{\pi}_{\gamma,\theta}^1 - \bar{\pi}_{\gamma,\theta}^2\|_{W^{1/2}}\right) \leqslant B_3\gamma\left|\kappa_1-\kappa_2\right| \ ,$$

where for any $i \in \{1,2\}$, $\theta \in \Theta$ *and* $\gamma \in (0,\bar{\gamma}]$, $\pi_{\gamma,\theta}^i$, *respectively* $\bar{\pi}_{\gamma,\theta}^i$, *is the invariant probability measure of* $\mathrm{S}_{i,\gamma,\theta}$, *respectively* $\bar{\mathrm{S}}_{i,\gamma,\theta}$, *given by* (22) *and associated with* $\kappa \leftarrow \kappa_i$. *In addition,* $W = W_m$ *with* $m \in \mathbb{N}^*$ *if* **H2** *holds and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta)$ *if* **H3** *holds.*

Proof. We only show that for any $\theta \in \Theta$, $\gamma \in (0,\bar{\gamma}]$ and $\kappa_i > 0$ with $i \in \{1,2\}$, $\|\pi_{\gamma,\theta}^1 - \pi_{\gamma,\theta}^2\|_{W^{1/2}} \leqslant B_3\gamma$ since the proof for $\bar{\pi}_{\gamma,\theta}^1$ and $\bar{\pi}_{\gamma,\theta}^2$ are similar. Let $\theta \in \Theta$, $\gamma \in (0,\bar{\gamma}]$, $x \in \mathbb{R}^d$ and $\kappa_i > 1/2$. Using Theorem 17 we have

$$\lim_{n\to+\infty} \|\delta_x \mathrm{S}_{1,\gamma,\theta}^n - \delta_x \mathrm{S}_{2,\gamma,\theta}^n\|_{W^{1/2}} = \|\pi_{1,\gamma,\theta} - \pi_{2,\gamma,\theta}\|_{W^{1/2}} \ .$$

Let $n = q\lceil 1/\gamma \rceil$. Using Theorem 17 with $a = 1/2$, that $W^{1/2}(x) \leqslant W(x)$ for any $x \in \mathbb{R}^d$, Lemma 20, Lemma 4 and Lemma 13 if **H2** holds or Lemma 14 if **H3** holds, we have

$$
\begin{aligned}
&\|\delta_x \mathrm{S}^n_{1,\gamma,\theta} - \delta_x \mathrm{S}^n_{2,\gamma,\theta}\|_{W^{1/2}} \leqslant \sum_{k=0}^{q-1} \|\delta_x \mathrm{S}^{(k+1)\lceil 1/\gamma \rceil}_{1,\gamma,\theta} \mathrm{S}^{(q-k-1)\lceil 1/\gamma \rceil}_{2,\gamma,\theta} - \delta_x \mathrm{S}^{k\lceil 1/\gamma \rceil}_{1,\gamma,\theta} \mathrm{S}^{(q-k)\lceil 1/\gamma \rceil}_{2,\gamma,\theta}\|_{W^{1/2}} \\
&\leqslant \sum_{k=0}^{q-1} A_{2,1/2} \rho_{1/2}^{q-k-1} \left\| \delta_x \mathrm{S}^{k\lceil 1/\gamma \rceil}_{1,\gamma,\theta} \left\{ \mathrm{S}^{\lceil 1/\gamma \rceil}_{1,\gamma,\theta} - \mathrm{S}^{\lceil 1/\gamma \rceil}_{2,\gamma,\theta} \right\} \right\|_{W^{1/2}} \\
&\leqslant A_{2,1/2} \sum_{k=0}^{q-1} \rho_{1/2}^{q-k-1} \bar{B}_3 \gamma \, |\kappa_1 - \kappa_2| \, \delta_x \mathrm{S}^{k\lceil 1/\gamma \rceil}_{1,\gamma,\theta} W(x) \\
&\leqslant A_{2,1/2} \sum_{k=0}^{q-1} \rho_{1/2}^{q-k-1} \bar{B}_3 \gamma \, |\kappa_1 - \kappa_2| \, (1 + b\lambda^{-\bar{\gamma}}/\log(1/\lambda)) W(x) \\
&\leqslant A_{2,1/2} \bar{B}_3 (1 + b\lambda^{-\bar{\gamma}}/\log(1/\lambda))/(1 - \rho_{1/2}) \, |\kappa_1 - \kappa_2| \, \gamma W(x) \, ,
\end{aligned}
$$

which concludes the proof with $B_3 = 2A_{2,1/2}\bar{B}_3(1+b\lambda^{-\bar{\gamma}}/\log(1/\lambda))/(1-\rho_{1/2})\kappa$ upon setting $x = 0$. □

Corollary 22. *Assume* **H1** *and* **H2** *or* **H3**. *Let* $\bar{\gamma} < 2/(\mathtt{m} + \mathtt{L})$ *if* **H2** *holds and* $\bar{\gamma} < 2/\mathtt{L}$ *if* **H3** *holds. Then for any* $\kappa > 0$, $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$, *we have*

$$
\max \left(\|\pi_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}}, \|\bar{\pi}_{\gamma,\theta} - \bar{\pi}_\theta\|_{W^{1/2}} \right) \leqslant \mathbf{\Psi}(\gamma) \, ,
$$

where for any $\gamma \in (0, \bar{\gamma}]$, $\pi_{\gamma,\theta}$ *is the invariant probability measure of* $\mathrm{S}_{\gamma,\theta}$ *given by* (22). *In addition,* $\mathbf{\Psi}(\gamma) = \tilde{\mathbf{\Psi}}(\gamma) + B_3\gamma|\kappa - 1|$, *where* $\tilde{\mathbf{\Psi}}$ *is given in Theorem 19 and* B_3 *in Proposition 21, and* $W = W_m$ *with* $m \in \mathbb{N}^*$ *if* **H2** *holds and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta)$ *if* **H3** *holds.*

Proof. We only show that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ we have $\|\pi_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} \leqslant \mathbf{\Psi}(\gamma)$ since the proof for $\bar{\pi}_{\gamma,\theta}$ and $\bar{\pi}_\theta$ are similar. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$. The proof is a direct application of Theorem 19 and Proposition 21 upon noticing that

$$
\|\pi_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} \leqslant \|\pi_{\gamma,\theta} - \pi^\sharp_{\gamma,\theta}\|_{W^{1/2}} + \|\pi^\sharp_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} \, ,
$$

where $\pi^\sharp_{\gamma,\theta}$ is the invariant probability measure of $\mathrm{S}_{\gamma,\theta}$ given by (22) and associated with $\kappa = 1$. □

Proposition 23. *Assume* **H1** *and* **H2** *or* **H3**. *Let* $\bar{\gamma} < 2/(\mathtt{m} + \mathtt{L})$ *if* **H2** *holds and* $\bar{\gamma} < 2/\mathtt{L}$ *if* **H3** *holds. Then there exists* $A_4 \geqslant 0$ *such that for any* $\kappa > 0$, $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ *with* $\gamma_2 < \gamma_1$, $a \in [1/4, 1/2]$ *and* $x \in \mathbb{R}^d$

$$
\begin{aligned}
&\max \left(\|\delta_x \mathrm{S}_{\gamma_1,\theta_1} - \delta_x \mathrm{S}_{\gamma_2,\theta_2}\|_{W^a}, \|\delta_x \bar{\mathrm{S}}_{\gamma_1,\theta_1} - \delta_x \bar{\mathrm{S}}_{\gamma_2,\theta_2}\|_{W^a} \right) \\
&\qquad \leqslant (\mathbf{\Lambda}(\gamma_1, \gamma_2) + \mathbf{\Lambda}(\gamma_1, \gamma_2) \, \|\theta_1 - \theta_2\|) W^{2a}(x) \, ,
\end{aligned}
$$

with

$$
\mathbf{\Lambda}_1(\gamma_1, \gamma_2) = A_4(\gamma_1/\gamma_2 - 1)^{1/2} \, , \qquad \mathbf{\Lambda}_2(\gamma_1, \gamma_2) = A_4 \gamma_2^{1/2} \, ,
$$

and where $W = W_m$ *with* $m \in \mathbb{N}$ *and* $m \geqslant 2$ *if* **H2** *is satisfied and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta)$ *if* **H3** *is satisfied.*

Proof. We only show that for any $\kappa > 0$, $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_2 < \gamma_1$, $a \in [1/4, 1/2]$ and $x \in \mathbb{R}^d$ we have $\|\delta_x \mathrm{S}_{\gamma_1,\theta_1} - \delta_x \mathrm{S}_{\gamma_2,\theta_2}\|_{W^a} \leqslant (\mathbf{\Lambda}(\gamma_1, \gamma_2) + \mathbf{\Lambda}(\gamma_1, \gamma_2) \, \|\theta_1 - \theta_2\|) W^{2a}(x)$ since the proof for $\bar{\mathrm{S}}_{\gamma_1,\theta_1}$ and $\bar{\mathrm{S}}_{\gamma_2,\theta_2}$ is similar. Let $a \in [1/4, 1/2]$, $\kappa > 0$, $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_2 < \gamma_1$. Using a generalized Pinsker inequality, see [32, Lemma 24], we have

$$
\begin{aligned}
&\|\delta_x \mathrm{S}_{\gamma_1,\theta_1} - \delta_x \mathrm{S}_{\gamma_2,\theta_2}\|_{W^a} \\
&\qquad \leqslant \sqrt{2} (\delta_x \mathrm{S}_{\gamma_1,\theta_1} W^{2a}(x) + \delta_x \mathrm{S}_{\gamma_2,\theta_2} W^{2a}(x))^{1/2} \mathrm{KL} \left(\delta_x \mathrm{S}_{\gamma_1,\theta_1} | \delta_x \mathrm{S}_{\gamma_2,\theta_2} \right)^{1/2} \, .
\end{aligned}
$$

Combining this result, Jensen's inequality and Lemma 13 if **H2** holds and Lemma 14 if **H3** holds, we obtain that

$$
\|\mathrm{S}_{\gamma_1,\theta_1} - \mathrm{S}_{\gamma_2,\theta_2}\|_{W^a} \leqslant 2(1 + b\bar{\gamma})^{1/2} \left\{ \mathrm{KL} \left(\delta_x \mathrm{S}_{\gamma_1,\theta_1} | \delta_x \mathrm{S}_{\gamma_2,\theta_2} \right) \right\}^{1/2} W^a(x) \, .
$$

Denote for $\upsilon \in \mathbb{R}^d$ and $\sigma > 0$, $\Upsilon_{\upsilon,\sigma}$ the d-dimensional Gaussian distribution with mean υ and covariance matrix $\sigma^2 \operatorname{Id}$. Let $\sigma_1, \sigma_2 > 0$ and $\upsilon_1, \upsilon_2 \in \mathbb{R}^d$, we have

$$\mathrm{KL}\left(\Upsilon_{\upsilon_1,\sigma_1 \operatorname{Id}} | \Upsilon_{\upsilon_2,\sigma_2 \operatorname{Id}}\right) = \left\{ \|\upsilon_1 - \upsilon_2\|^2 + d(\sigma_1^2 - \sigma_2^2) \right\} \Big/ (2\sigma_2^2) + d \log(\sigma_2/\sigma_1) \ .$$

Since $\gamma_2 < \gamma_1$, we have

$$\begin{aligned} &\mathrm{KL}\left(\delta_x \mathrm{S}_{\gamma_1,\theta_1} | \delta_x \mathrm{S}_{\gamma_2,\theta_2}\right) \\ &\qquad \leqslant d(\gamma_1/\gamma_2 - 1)/2 + \left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_1\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_2,\theta_2}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_1}}(x)) \right\|^2 \Big/ (4\gamma_2) \ , \end{aligned} \tag{38}$$

with $\mathcal{T}_{\gamma,\theta}(z) = z - \gamma \nabla_x V_\theta(z)$ for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. We have

$$\begin{aligned} &(1/4) \left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_1\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_2,\theta_2}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\|^2 \\ &\leqslant \left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_1\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_1}}(x)) \right\|^2 + \left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\|^2 \\ &+ \left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) - \mathcal{T}_{\gamma_2,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\|^2 + \left\| \mathcal{T}_{\gamma_2,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) - \mathcal{T}_{\gamma_2,\theta_2}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\|^2 . \end{aligned} \tag{39}$$

First using **H1**, [61, Theorem 2.1.5, Equation (2.1.8)] and Lemma 7 we have

$$\begin{aligned} &\left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_1\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_1}}(x)) \right\| \\ &\qquad \leqslant \left\| \mathrm{prox}^{\gamma_1\kappa}_{U_{\theta_1}}(x) - \mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_1}}(x) \right\| \leqslant 2\mathtt{M} \left| \gamma_1\kappa - \gamma_2\kappa \right| \ . \end{aligned} \tag{40}$$

Second we have using (15), **H1**, [61, Theorem 2.1.5, Equation (2.1.8)] and **H4**

$$\begin{aligned} &\left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\| \\ &\qquad \leqslant \gamma_2\kappa \left\| \nabla_x U^{\gamma_2\kappa}_{\theta_1}(x) - \nabla_x U^{\gamma_2\kappa}_{\theta_2}(x) \right\| \leqslant \sup_{t \in [0,\bar{\gamma}\kappa]} \{\mathtt{f}_\theta(t)\} \gamma_2\kappa \left\| \theta_1 - \theta_2 \right\| (1 + \|x\|) \ . \end{aligned} \tag{41}$$

Third using **H1** and Lemma 5 we have that

$$\begin{aligned} \left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) - \mathcal{T}_{\gamma_2,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\| &\leqslant (\gamma_1 - \gamma_2) \left\| \nabla_x V_{\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\| \\ &\leqslant (\gamma_1 - \gamma_2)\mathrm{L} \left\| \mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x) - x^\star_{\theta_1} \right\| \\ &\leqslant (\gamma_1 - \gamma_2)\mathrm{L}(R_{V,1} + \bar{\gamma}\kappa\mathtt{M} + \|x\|) \ . \end{aligned} \tag{42}$$

Finally using **H1**, **H4** and Lemma 5 we have that

$$\begin{aligned} &\left\| \mathcal{T}_{\gamma_2,\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) - \mathcal{T}_{\gamma_2,\theta_2}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\| \\ &\quad \leqslant \gamma_2 \left\| \nabla_x V_{\theta_1}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) - \nabla_x V_{\theta_2}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\| \\ &\quad \leqslant \gamma_2\mathtt{M}_\Theta \left\| \theta_1 - \theta_2 \right\| (1 + \| \mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x) \|) \leqslant \gamma_2\mathtt{M}_\Theta \left\| \theta_1 - \theta_2 \right\| (1 + \bar{\gamma}\kappa\mathtt{M} + \|x\|) \ . \end{aligned} \tag{43}$$

Therefore, combining (40), (41), (42) and (43) in (39), there exists $A_{4,1} \geqslant 0$ such that for any $\gamma_1, \gamma_2 > 0$ with $\gamma_2 < \gamma_1$ and $\theta_1, \theta_2 \in \Theta$

$$\left\| \mathcal{T}_{\gamma_1,\theta_1}(\mathrm{prox}^{\gamma_1\kappa}_{U_{\theta_1}}(x)) - \mathcal{T}_{\gamma_2,\theta_2}(\mathrm{prox}^{\gamma_2\kappa}_{U_{\theta_2}}(x)) \right\|^2 \leqslant A_{4,1} \left[(\gamma_1 - \gamma_2)^2 + \gamma_2^2 \left\| \theta_1 - \theta_2 \right\|^2 \right] W^{2a}(x) \ .$$

Using this result in (38), there exists $A_{4,2} \geqslant 0$ such that

$$\mathrm{KL}\left(\delta_x \mathrm{S}_{\gamma_1,\theta_1} | \delta_x \mathrm{S}_{\gamma_2,\theta_2}\right) \leqslant A_{4,2} \left[(\gamma_1/\gamma_2 - 1) + \gamma_2 \left\| \theta_1 - \theta_2 \right\|^2 \right] W^{2a}(x) \ ,$$

which implies the announced result upon setting $A_4 = 2\sqrt{A_{4,2}}(1 + b\bar{\gamma})^{1/2}$ and using that for any $u, v \geqslant 0$, $\sqrt{u+v} \leqslant \sqrt{u} + \sqrt{v}$. □

D.3 Checking [80, H1, H2] for MYULA

In this Section, similarly to Appendix D.3 for PULA, we show that [80, H1, H2] hold for MYULA.

Lemma 24. *Assume* **H1** *and* **H2**. *Then for any* $\kappa > 0$, $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$, $\mathrm{R}_{\gamma,\theta}$ *and* $\bar{\mathrm{R}}_{\gamma,\theta}$ *satisfy* $\mathbf{D}_{\mathrm{d}}(W_2, \lambda_2^\gamma, b_2\gamma)$ *with*

$$\begin{aligned}
&\lambda_2 = \exp\left[-\boldsymbol{\kappa}/2\right] \,, \\
&b_2 = \left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2 + 2\bar{\gamma}\mathtt{M}\mathtt{L}R_{V,1} + \bar{\gamma}\mathtt{M}^2 + 2d + 2\mathtt{M}^2(1+\bar{\gamma}\mathtt{L})^2\boldsymbol{\kappa}^{-1} + \boldsymbol{\kappa}/2 \,, \\
&\boldsymbol{\kappa} = \mathtt{m}\mathtt{L}/(\mathtt{m}+\mathtt{L}) \,.
\end{aligned}$$

In addition, for any $m \in \mathbb{N}^*$, *there exist* $\lambda_m \in (0,1)$, $b_m \geqslant 0$ *such that for any* $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < 2/(\mathtt{m}+\mathtt{L})$, $\mathrm{R}_{\gamma,\theta}$ *and* $\bar{\mathrm{R}}_{\gamma,\theta}$ *satisfy* $\mathbf{D}_{\mathrm{d}}(W_m, \lambda_m^\gamma, b_m\gamma)$.

Proof. We show the property for $\mathrm{R}_{\gamma,\theta}$ only as the proof for $\bar{\mathrm{R}}_{\gamma,\theta}$ is identical. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Let Z be a d-dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 11 we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \|y\|^2 \mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y) &= \mathbb{E}\left[\left\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\boldsymbol{\kappa}}(x) + \sqrt{2\gamma}Z\right\|^2\right] \\
&= \|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\boldsymbol{\kappa}}(x)\| + 2\gamma d \\
&\leqslant (1 - \gamma\boldsymbol{\kappa}/2)\|x\|^2 + \gamma\left[\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2\right. \\
&\qquad \left. + 2\bar{\gamma}\mathtt{M}\mathtt{L}R_{V,1} + \bar{\gamma}\mathtt{M}^2 + 2d + 2\mathtt{M}^2(1+\bar{\gamma}\mathtt{L})^2\boldsymbol{\kappa}^{-1}\right] \,.
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\int_{\mathbb{R}^d} (1 + \|y\|^2)\mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y) \leqslant (1 - \gamma\boldsymbol{\kappa}/2)(1 + \|x\|^2) + \gamma\left[\left\{(2/(\mathtt{m}+\mathtt{L}) - \bar{\gamma})^{-1} + 4\boldsymbol{\kappa}\right\} R_{V,1}^2\right. \\
\left. + 2\bar{\gamma}\mathtt{M}\mathtt{L}R_{V,1} + \bar{\gamma}\mathtt{M}^2 + 2d + 2\mathtt{M}^2(1+\bar{\gamma}\mathtt{L})^2\boldsymbol{\kappa}^{-1} + \boldsymbol{\kappa}/2\right] \,,
\end{aligned}$$

which concludes the first part of the proof. The proof of the result for $W = W_m$ with $m \in \mathbb{N}^*$ is a straightforward adaptation of the one of Lemma 13 and is left to the reader. □

Lemma 25. *Assume* **H1** *and* **H3**. *Then for any* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(2/\mathtt{L}, \eta/(2\mathtt{M}\mathtt{L}))$, $\kappa > 0$ *and* $\theta \in \Theta$, $\mathrm{R}_{\gamma,\theta}$ *and* $\bar{\mathrm{R}}_{\gamma,\theta}$ *satisfy* $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, b\gamma)$ *with*

$$\begin{aligned}
&\lambda = \mathrm{e}^{-\alpha^2} \,, \\
&b_e = (4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + (3/2)\bar{\gamma}\mathtt{M}^2 + \mathtt{c} + \mathtt{M}R_{U,1} + R_{U,2} + \bar{\gamma}\mathtt{M}\mathtt{L}R_{V,2} + d + 2\alpha \,, \\
&b = \alpha b_e \mathrm{e}^{\alpha\bar{\gamma}b_e} W(R) \,, \\
&W = W_\alpha \,, \qquad \alpha < \eta/8 \,, \\
&R_\eta = \max\left(2b_e/(\eta - 8\alpha), 1\right) \,.
\end{aligned} \tag{44}$$

Proof. We show the property for $\mathrm{R}_{\gamma,\theta}$ only as the proof for $\bar{\mathrm{R}}_{\gamma,\theta}$ is identical. Let $\kappa > 0$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and Z be a d-dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 12 we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|y\|^2 \mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y) = \|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\boldsymbol{\kappa}}\|^2 + 2\gamma d \\
&\leqslant \|x\|^2 + \gamma\left[(2/L - \bar{\gamma})^{-1} R_{V,1}^2 + 3\bar{\gamma}\mathtt{M}^2 + 2\mathtt{c} + 2(\mathtt{M}R_{U,1} + R_{U,2}) + 2\bar{\gamma}\mathtt{M}\mathtt{L}R_{V,2} + 2d - \eta\|x\|\right] \,.
\end{aligned}$$

Using the log-Sobolev inequality [9, Proposition 5.4.1] and Jensen's inequality we get that

$$\begin{aligned}
\mathrm{R}_{\gamma,\theta} W(x) &\leqslant \exp\left[\alpha \mathrm{R}_{\gamma,\theta}\phi(x) + \alpha^2\gamma\right] \\
&\leqslant \exp\left[\alpha\left(1 + \int_{\mathbb{R}^d} \|y\|^2 \mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y)\right)^{1/2} + \alpha^2\gamma\right] \,.
\end{aligned} \tag{45}$$

We now distinguish two cases:

(a) If $\|x\| \geqslant R_\eta$, recalling that R_η is given in (44), then

$$(2/L - \bar{\gamma})^{-1} R_{V,1}^2 + 3\bar{\gamma}\mathtt{M}^2 + 2\mathtt{c} + 2(\mathtt{M}R_{U,1} + R_{U,2}) + 2\bar{\gamma}\mathtt{ML}R_{V,2} + 2d - \eta\|x\| \leqslant -8\alpha\|x\| \ .$$

In this case using that $\phi^{-1}(x)\|x\| \geqslant 1/2$ and that for any $t \geqslant 0$, $\sqrt{1+t} \leqslant 1 + t/2$ we have

$$\begin{aligned}
&\left(1 + \int_{\mathbb{R}^d} \|y\|^2 \, \mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y)\right)^{1/2} - \phi(x) \leqslant \\
&\quad \leqslant \gamma\phi^{-1}(x)\left((2/L - \bar{\gamma})^{-1} R_{V,1}^2 + 3\bar{\gamma}\mathtt{M}^2 + 2\mathtt{c} + 2(\mathtt{M}R_{U,1} + R_{U,2}) + 2\bar{\gamma}\mathtt{ML}R_{V,2} + 2d - \eta\|x\|\right)/2 \\
&\quad \leqslant -4\alpha\gamma\phi^{-1}(x)\|x\| \leqslant -2\alpha\gamma \ .
\end{aligned}$$

Hence,

$$\mathrm{R}_{\gamma,\theta} W(x) \leqslant \left[\alpha\left(1 + \int_{\mathbb{R}^d} \|y\|^2 \, \mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y)\right)^{1/2} + \alpha^2\gamma\right] \leqslant \mathrm{e}^{-\alpha^2\gamma} W(x) \ .$$

(b) If $\|x\| \leqslant R_\eta$ then using that for any $t \geqslant 0$, $\sqrt{1+t} \leqslant 1 + t/2$ we have

$$\begin{aligned}
&\left(1 + \int_{\mathbb{R}^d} \|y\|^2 \, \mathrm{R}_{\gamma,\theta}(x, \mathrm{d}y)\right)^{1/2} - \phi(x) \\
&\qquad \leqslant \gamma((4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + (3/2)\bar{\gamma}\mathtt{M}^2 + \mathtt{c} + \mathtt{M}R_{U,1} + R_{U,2} + \bar{\gamma}\mathtt{ML}R_{V,2} + d) \ .
\end{aligned}$$

Therefore, using (45), we get

$$\begin{aligned}
&\mathrm{R}_{\gamma,\theta} W(x) \\
&\qquad \leqslant \exp\left[\alpha\gamma\left\{(4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + (3/2)\bar{\gamma}\mathtt{M}^2 + \mathtt{c} + \mathtt{M}R_{U,1} + R_{U,2} + \bar{\gamma}\mathtt{ML}R_{V,2} + d + \alpha\right\}\right] W(x) \ .
\end{aligned}$$

Since for all $a \geqslant b$, $\mathrm{e}^a - \mathrm{e}^b \leqslant (a-b)\mathrm{e}^a$ we obtain that

$$\mathrm{R}_{\gamma,\theta} W(x) \leqslant \lambda^\gamma W(x) + \gamma\alpha b_e \mathrm{e}^{\alpha\bar{\gamma}b_e} W(R_\eta) \ ,$$

which concludes the proof.

□

Lemma 26. *Assume* **H1**, **H2** *or* **H3**, *and let* $(X_k^n, \bar{X}_k^n)_{n\in\mathbb{N}, k\in\{0,\ldots,m_n\}}$ *be given by* (23) *with* $\{(\mathrm{K}_{\gamma,\theta}, \bar{\mathrm{K}}_{\gamma,\theta}) : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\} = \{(\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta}) : \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$ *and* $\kappa > 0$. *Then there exists* $\bar{A}_1 \geqslant 1$ *such that for any* $n, p \in \mathbb{N}$ *and* $k \in \{0, \ldots, m_n\}$

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{R}_{\gamma_n,\theta_n}^p W(X_k^n) \,\middle|\, X_0^0\right] \leqslant \bar{A}_1 W(X_0^0) \ , \\
&\mathbb{E}\left[\bar{\mathrm{R}}_{\gamma_n,\theta_n}^p W(\bar{X}_k^n) \,\middle|\, \bar{X}_0^0\right] \leqslant \bar{A}_1 W(\bar{X}_0^0) \ , \\
&\mathbb{E}\left[W(X_0^0)\right] < +\infty \ , \qquad \mathbb{E}\left[W(\bar{X}_0^0)\right] < +\infty \ .
\end{aligned}$$

with $W = W_m$ *with* $m \in \mathbb{N}^*$ *and* $\bar{\gamma} < 2/(\mathtt{m} + \mathtt{L})$ *if* **H2** *holds and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta/8)$ *and* $\bar{\gamma} < \min\{2/\mathtt{L}, \eta/(2\mathtt{ML})\}$ *if* **H3** *holds.*

Proof. Combining [80, Lemma S15] and Lemma 24 if **H2** holds or Lemma 25 if **H3** holds conclude the proof. □

Lemma 27. *For any* $\kappa > 1/2$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < (2 - 1/\kappa)/\mathtt{L}$ *and* $x, y \in \mathbb{R}^d$

$$\max\left(\|\delta_x \mathrm{R}_{\gamma,\theta}^{\lceil 1/\gamma\rceil} - \delta_y \mathrm{R}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\|_{\mathrm{TV}}, \|\delta_x \bar{\mathrm{R}}_{\gamma,\theta}^{\lceil 1/\gamma\rceil} - \delta_y \bar{\mathrm{R}}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\|_{\mathrm{TV}}\right) \leqslant 1 - 2\mathbf{\Phi}\left\{-\|x-y\|/(2\sqrt{2})\right\} \ ,$$

where $\mathbf{\Phi}$ *is the cumulative distribution function of the standard normal distribution on* $\mathbb{R}$.

Proof. We only show that for any $\kappa > 1/2$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < (2 - 1/\kappa)/\mathtt{L}$ and $x, y \in \mathbb{R}^d$, we have $\|\delta_x \mathrm{R}_{\gamma,\theta}^{\lceil 1/\gamma \rceil} - \delta_y \mathrm{R}_{\gamma,\theta}^{\lceil 1/\gamma \rceil}\|_{\mathrm{TV}} \leqslant 1 - 2\mathbf{\Phi}\left\{-\|x-y\|/(2\sqrt{2})\right\}$ as the proof of for $\bar{\mathrm{R}}_{\gamma,\theta}$ is similar. Let $\kappa > 1/2$, $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$. We have that $x \mapsto V_\theta(x) + U_\theta^{\gamma\kappa}(x)$ is convex, continuously differentiable and satisfies for any $x, y \in \mathbb{R}^d$

$$\|\nabla_x V_\theta(x) + \nabla_x U_\theta^{\gamma\kappa}(x) - \nabla_x V_\theta(y) - \nabla_x U_\theta^{\gamma\kappa}(y)\| \leqslant \{\mathtt{L} + 1/(\gamma\kappa)\}\|x - y\| \ ,$$

Combining this result with [61, Theorem 2.1.5, Equation (2.1.8)] and the fact that $\gamma \leqslant 2/\{\mathtt{L} + 1/(\gamma\kappa)\}$ since $\bar{\gamma} \leqslant (2 - 1/\kappa)/\mathtt{L}$, we have for any $x, y \in \mathbb{R}^d$

$$\|x - \gamma\nabla_x V_\theta(x) - \gamma\nabla_x U_\theta^{\gamma\kappa}(x) + y + \gamma\nabla_x V_\theta(y) + \gamma\nabla_x U_\theta^{\gamma\kappa}(y)\| \leqslant \|x - y\| \ .$$

The proof is then an application of [81, Proposition 3b] with $\ell \leftarrow 1$. □

Theorem 28. *Assume* **H1** *and* **H2** *or* **H3**. *Let $\kappa > 1/2$, $\bar{\gamma} < \min\{(2 - 1/\kappa)/\mathtt{L}, 2/(\mathtt{m} + \mathtt{L})\}$ if* **H2** *holds and $\bar{\gamma} < \min\{(2 - 1/\kappa)/\mathtt{L}, \eta/(2\mathtt{ML})\}$ if* **H3** *holds. Then for any $a \in (0, 1]$, there exist $\bar{A}_{2,a} \geqslant 0$ and $\rho_a \in (0, 1)$ such that for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$, $\mathrm{R}_{\gamma,\theta}$ and $\bar{\mathrm{R}}_{\gamma,\theta}$ admit invariant probability measures $\pi_{\gamma,\theta}$, respectively $\bar{\pi}_{\gamma,\theta}$. In addition, for any $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$ we have*

$$\max\left(\|\delta_x \mathrm{R}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_{W^a}, \|\delta_x \bar{\mathrm{R}}_{\gamma,\theta}^n - \bar{\pi}_{\gamma,\theta}\|_{W^a}\right) \leqslant \bar{A}_{2,a}\bar{\rho}_a^{\gamma n} W^a(x) \ ,$$
$$\max\left(\|\delta_x \mathrm{R}_{\gamma,\theta}^n - \delta_y \mathrm{R}_{\gamma,\theta}^n\|_{W^a}, \|\delta_x \bar{\mathrm{R}}_{\gamma,\theta}^n - \delta_y \bar{\mathrm{R}}_{\gamma,\theta}^n\|_{W^a}\right) \leqslant \bar{A}_{2,a}\bar{\rho}_a^{\gamma n}\left\{W^a(x) + W^a(y)\right\} \ ,$$

with $W = W_m$ and $m \in \mathbb{N}^*$ if **H2** *holds and $W = W_\alpha$ with $\alpha < \min(\kappa\eta/4, \eta/8)$ if* **H3** *holds.*

Proof. The proof is similar to the one of Theorem 17. □

Proposition 29. *Assume* **H1** *and* **H2** *or* **H3**. *Let $\kappa > 1/2$ and $\bar{\gamma} < \min\{(2 - 1/\kappa)/\mathtt{L}, 2/(\mathtt{m} + \mathtt{L})\}$ if* **H2** *holds and $\bar{\gamma} < \min\{(2 - 1/\kappa)/\mathtt{L}, \eta/(2\mathtt{ML})\}$ if* **H3** *holds. Then there exists $\bar{B}_{3,1} \geqslant 0$ such that for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $\kappa_i \geqslant \kappa$ with $i \in \{1, 2\}$ we have*

$$\max\left(\|\pi_{\gamma,\theta}^1 - \pi_{\gamma,\theta}^2\|_{W^{1/2}}, \|\bar{\pi}_{\gamma,\theta}^1 - \bar{\pi}_{\gamma,\theta}^2\|_{W^{1/2}}\right) \leqslant \bar{B}_{3,1}\gamma \ ,$$

where for any $i \in \{1, 2\}$, $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, $\pi_{\gamma,\theta}^i$, respectively $\bar{\pi}_{\gamma,\theta}^i$, is the invariant probability measure of $\mathrm{R}_{i,\gamma,\theta}$, respectively $\bar{\mathrm{R}}_{i,\gamma,\theta}$, given by (21) and associated with $\kappa \leftarrow \kappa_i$. In addition, $W = W_m$ with $m \in \mathbb{N}^*$ if **H2** *holds and $W = W_\alpha$ with $\alpha < \min(\kappa\eta/4, \eta/8)$ if* **H3** *holds.*

Proof. The proof is similar to the one of Proposition 21 upon setting for any $i \in \{1, 2\}$ and $(\omega_s)_{s\in[0,T]} \in \mathrm{C}([0, T], \mathbb{R}^d)$ with $T = \gamma\lceil 1/\gamma \rceil$

$$b_i(t, (\omega_s)_{s\in[0,T]}) = \omega_{\lfloor t/\gamma \rfloor \gamma} - \gamma\nabla_x V_\theta(\omega_{\lfloor t/\gamma \rfloor \gamma}) - \gamma\nabla_x U_\theta^{\gamma\kappa_i(\gamma)}(\omega_{\lfloor t/\gamma \rfloor \gamma}) \ ,$$

and replacing (36) in Lemma 20 by

$$\left\|b_1(t, (\omega_s)_{s\in[0,T]}) - b_2(t, (\omega_s)_{s\in[0,T]})\right\|^2 = \left\|-\gamma\nabla_x U_\theta^{\gamma\kappa_1}(\omega_{\lfloor t/\gamma \rfloor \gamma}) + \gamma\nabla_x U_\theta^{\gamma\kappa_2}(\omega_{\lfloor t/\gamma \rfloor \gamma})\right\|^2 \leqslant 4\gamma^2\mathtt{M}^2 \ .$$

□

Proposition 30. *Assume* **H1** *and* **H2** *or* **H3**. *Let $\kappa > 1/2$ and $\bar{\gamma} < \min\{(2 - 1/\kappa)/\mathtt{L}, 2/(\mathtt{m} + \mathtt{L}), \mathtt{L}^{-1}\}$ if* **H2** *holds and $\bar{\gamma} < \min\{(2 - 1/\kappa)/\mathtt{L}, \eta/(2\mathtt{ML}), \mathtt{L}^{-1}\}$ if* **H3** *holds. Then there exists $\bar{B}_{3,2} \geqslant 0$ such that for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $\kappa_i > 1/2$ with $i \in \{1, 2\}$ we have*

$$\max\left(\|\pi_{\gamma,\theta}^\flat - \pi_{\gamma,\theta}^\sharp\|_{W^{1/2}}, \|\bar{\pi}_{\gamma,\theta}^\flat - \bar{\pi}_{\gamma,\theta}^\sharp\|_{W^{1/2}}\right) \leqslant \bar{B}_{3,2}\gamma^2 \ ,$$

where for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, $\pi_{\gamma,\theta}^\flat$, respectively $\bar{\pi}_{\gamma,\theta}^\flat$, is the invariant probability measure of $\mathrm{R}_{\gamma,\theta}$, respectively $\bar{\mathrm{R}}_{\gamma,\theta}$, given by (21) and associated with $\kappa = 1$ and $\pi_{\gamma,\theta}^\sharp$, respectively $\bar{\pi}_{\gamma,\theta}^\sharp$, is the invariant probability measure of $\mathrm{S}_{\gamma,\theta}$, respectively $\bar{\mathrm{S}}_{\gamma,\theta}$, given by (22) and associated with $\kappa = 1$. In addition, $W = W_m$ with $m \in \mathbb{N}^*$ if **H2** *holds and $W = W_\alpha$ with $\alpha < \min(\kappa\eta/4, \eta/8)$ if* **H3** *holds.*

Proof. The proof is similar to the one of Proposition 21 upon setting for any $(\omega_s)_{s\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$ with $T = \gamma\lceil 1/\gamma\rceil$

$$b_1(t,(\omega_s)_{s\in[0,T]}) = \mathrm{prox}^{\gamma}_{U_\theta}(\omega_{\lfloor t/\gamma\rfloor\gamma}) - \gamma\nabla_x V_\theta(\mathrm{prox}^{\gamma}_{U_\theta}(\omega_{\lfloor t/\gamma\rfloor\gamma}))\;,$$
$$b_2(t,(\omega_s)_{s\in[0,T]}) = \omega_{\lfloor t/\gamma\rfloor\gamma} - \gamma\nabla_x V_\theta(\omega_{\lfloor t/\gamma\rfloor\gamma}) - \gamma\nabla_x U^{\gamma}_\theta(\omega_{\lfloor t/\gamma\rfloor\gamma})\;,$$

and replacing (36) in Lemma 20 and using (15) and Lemma 5 we get

$$\begin{aligned}
&\left\|b_1(t,(\omega_s)_{s\in[0,T]}) - b_2(t,(\omega_s)_{s\in[0,T]})\right\|^2 \\
&= \|\,\mathrm{prox}^{\gamma}_{U_\theta}(\omega_{\lfloor t/\gamma\rfloor\gamma})) - \gamma\nabla_x V_\theta(\mathrm{prox}^{\gamma}_{U_\theta}(\omega_{\lfloor t/\gamma\rfloor\gamma})) - \omega_{\lfloor t/\gamma\rfloor\gamma} \\
&\qquad + \gamma\nabla_x V_\theta(\omega_{\lfloor t/\gamma\rfloor\gamma})) + \gamma(\omega_{\lfloor t/\gamma\rfloor\gamma} - \mathrm{prox}^{\gamma}_{U_\theta}(\omega_{\lfloor t/\gamma\rfloor\gamma}))/\gamma\|^2 \\
&= \gamma^2\left\|\nabla_x V_\theta(\mathrm{prox}^{\gamma}_{U_\theta}(\omega_{\lfloor t/\gamma\rfloor\gamma}))) - \nabla_x V_\theta(\omega_{\lfloor t/\gamma\rfloor\gamma}))\right\|^2 \leqslant \mathtt{L}^2\mathtt{M}^2\gamma^4\;.
\end{aligned}$$

□

Proposition 31. *Assume* H1 *and* H2 *or* H3. *Let* $\kappa > 1/2$, $\bar\gamma < \min\{(2-1/\kappa)/\mathtt{L}, 2/(\mathtt{m}+\mathtt{L}), \mathtt{L}^{-1}\}$ *if* H2 *holds and* $\bar\gamma < \min\{(2-1/\kappa)/\mathtt{L}, \eta/(2\mathtt{ML}), \mathtt{L}^{-1}\}$ *if* H3 *holds. Then for any* $\theta\in\Theta$ *and* $\gamma\in(0,\bar\gamma]$, *we have*

$$\max\left(\|\pi_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}}, \|\bar\pi_{\gamma,\theta} - \bar\pi_\theta\|_{W^{1/2}}\right) \leqslant \bar{\mathbf{\Psi}}(\gamma)\;,$$

where for any $i\in\{1,2\}$, $\theta\in\Theta$ *and* $\gamma\in(0,\bar\gamma]$, $\pi^i_{\gamma,\theta}$, *respectively* $\bar\pi^i_{\gamma,\theta}$, *is the invariant probability measure of* $\mathrm{R}_{i,\gamma,\theta}$, *respectively* $\bar{\mathrm{R}}_{i,\gamma,\theta}$, *given by* (21) *and associated with* $\kappa \leftarrow \kappa_i$. *In addition,* $\bar{\mathbf{\Psi}}(\gamma) = \tilde{\mathbf{\Psi}}(\gamma) + \bar B_{3,1}\gamma + \bar B_{3,2}\gamma^2$, *where* $\tilde{\mathbf{\Psi}}$ *is given in Theorem 19 and* B_3 *in Proposition 21, and* $W = W_m$ *with* $m\in\mathbb{N}^*$ *if* H2 *holds and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta/8)$ *if* H3 *holds.*

Proof. We only show that for any $\theta\in\Theta$ and $\gamma\in(0,\bar\gamma]$, $\|\pi_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} \leqslant \bar{\mathbf{\Psi}}(\gamma)$ as the proof for $\bar\pi_{\gamma,\theta}$ and $\bar\pi_\theta$ is similar. First note that for any $\theta\in\Theta$ and $\gamma\in(0,\bar\gamma]$ we have

$$\|\pi_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}} \leqslant \|\pi_{\gamma,\theta} - \pi^\flat_{\gamma,\theta}\|_{W^{1/2}} + \|\pi^\flat_{\gamma,\theta} - \pi^\sharp_{\gamma,\theta}\|_{W^{1/2}} + \|\pi^\sharp_{\gamma,\theta} - \pi_\theta\|_{W^{1/2}}\;,$$

where for any $\theta\in\Theta$ and $\gamma\in(0,\bar\gamma]$, $\pi^\flat_{\gamma,\theta}$ is the invariant probability measure of $\mathrm{R}_{\gamma,\theta}$ given by (21) and associated with $\kappa = 1$ and $\pi^\sharp_{\gamma,\theta}$ is the invariant probability measure of $\mathrm{S}_{\gamma,\theta}$ and associated with $\kappa = 1$. We conclude the proof upon combining Proposition 29, Proposition 30 and Theorem 19. □

Proposition 32. *Assume* H1 *and* H2 *or* H3. *Let* $\kappa > 0$ *and* $\bar\gamma < \min\{(2-1/\kappa)/\mathtt{L}, 2/(\mathtt{m}+\mathtt{L})\}$ *if* H2 *holds and* $\bar\gamma < \min\{(2-1/\kappa)/\mathtt{L}, \eta/(2\mathtt{ML})\}$ *if* H3 *holds. Then there exists* $\bar A_4 \geqslant 0$ *such that for any* $\theta_1,\theta_2\in\Theta$, $\gamma_1,\gamma_2\in(0,\bar\gamma]$ *with* $\gamma_2 < \gamma_1$, $a\in[1/4,1/2]$ *and* $x\in\mathbb{R}^d$

$$\begin{aligned}
&\max\left(\|\delta_x\mathrm{R}_{\gamma_1,\theta_1} - \delta_x\mathrm{R}_{\gamma_2,\theta_2}\|_{W^a}, \|\delta_x\bar{\mathrm{R}}_{\gamma_1,\theta_1} - \delta_x\bar{\mathrm{R}}_{\gamma_2,\theta_2}\|_{W^a}\right) \\
&\qquad\qquad \leqslant (\bar{\mathbf{\Lambda}}_1(\gamma_1,\gamma_2) + \bar{\mathbf{\Lambda}}_2(\gamma_1,\gamma_2)\,\|\theta_1-\theta_2\|)W^{2a}(x)\;,
\end{aligned}$$

with

$$\bar{\mathbf{\Lambda}}_1(\gamma_1,\gamma_2) = \bar A_4(\gamma_1/\gamma_2 - 1)^{1/2}\;, \qquad \bar{\mathbf{\Lambda}}_2(\gamma_1,\gamma_2) = \bar A_4\gamma_2^{1/2}\;,$$

and where $W = W_m$ *with* $m\in\mathbb{N}$ *and* $m\geqslant 2$ *if* H2 *is satisfied and* $W = W_\alpha$ *with* $\alpha < \min(\kappa\eta/4, \eta/8)$ *if* H3 *is satisfied.*

Proof. We only show that for any $\theta_1,\theta_2\in\Theta$, $\gamma_1,\gamma_2\in(0,\bar\gamma]$ with $\gamma_2<\gamma_1$, $a\in[1/4,1/2]$ and $x\in\mathbb{R}^d$, we have $\|\delta_x\mathrm{R}_{\gamma_1,\theta_1} - \delta_x\mathrm{R}_{\gamma_2,\theta_2}\|_{W^a} \leqslant (\bar{\mathbf{\Lambda}}(\gamma_1,\gamma_2) + \bar{\mathbf{\Lambda}}(\gamma_1,\gamma_2)\,\|\theta_1-\theta_2\|)W^{2a}(x)$ since the proof $\bar{\mathrm{R}}_{\gamma_1,\theta_1}$ and $\bar{\mathrm{R}}_{\gamma_2,\theta_2}$ is similar. Let $a\in[1/4,1/2]$, $\theta_1,\theta_2\in\Theta$, $\gamma_1,\gamma_2\in(0,\bar\gamma]$ with $\gamma_2<\gamma_1$. Using a generalized Pinsker inequality [32, Lemma 24] we have

$$\begin{aligned}
&\|\delta_x\mathrm{R}_{\gamma_1,\theta_1} - \delta_x\mathrm{R}_{\gamma_2,\theta_2}\|_{W^a} \\
&\qquad\leqslant \sqrt{2}(\delta_x\mathrm{R}_{\gamma_1,\theta_1}W^{2a}(x) + \delta_x\mathrm{R}_{\gamma_2,\theta_2}W^{2a}(x))^{1/2}\mathrm{KL}\left(\delta_x\mathrm{R}_{\gamma_1,\theta_1}|\delta_x\mathrm{R}_{\gamma_2,\theta_2}\right)^{1/2}\;.
\end{aligned}$$

Combining this result, Jensen's inequality and Lemma 13 if H2 holds and Lemma 14 if H3 holds, we obtain that

$$\|\delta_x\mathrm{R}_{\gamma_1,\theta_1} - \delta_x\mathrm{R}_{\gamma_2,\theta_2}\|_{W^a} \leqslant 2(1+b\bar\gamma)^{1/2}\mathrm{KL}\left(\delta_x\mathrm{R}_{\gamma_1,\theta_1}|\delta_x\mathrm{R}_{\gamma_2,\theta_2}\right)^{1/2}W^a(x)\;.$$

Denote for $\upsilon \in \mathbb{R}^d$ and $\mathfrak{s} > 0$, $\Upsilon_{\upsilon,\mathfrak{s}}$ the d-dimensional Gaussian distribution with mean υ and covariance matrix $\mathfrak{s}^2 \operatorname{Id}$. Let $\mathfrak{s}_1, \mathfrak{s}_2 > 0$ and $\upsilon_1, \upsilon_2 \in \mathbb{R}^d$, we have

$$\mathrm{KL}\left(\Upsilon_{\upsilon_1,\mathfrak{s}_1 \operatorname{Id}} | \Upsilon_{\upsilon_2,\mathfrak{s}_2 \operatorname{Id}}\right) = \left\{ \|\upsilon_1 - \upsilon_2\|^2 + d(\mathfrak{s}_1^2 - \mathfrak{s}_2^2) \right\} / (2\mathfrak{s}_2^2) + d \log(\mathfrak{s}_2/\mathfrak{s}_1) \; .$$

Since $\gamma_2 < \gamma_1$, we have

$$\begin{aligned} &\mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right) \\ &\quad \leqslant d(\gamma_1/\gamma_2 - 1)/2 + \|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_1 \nabla_x V_{\theta_1}(x) + \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x) - \gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x)\|^2 / (4\gamma_2) \; , \end{aligned} \tag{46}$$

We have

$$\begin{aligned} &\|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_1 \nabla_x V_{\theta_1}(x) + \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x) - \gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x)\|^2 \\ &\qquad \leqslant 4 \|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_2 \nabla_x V_{\theta_1}(x)\|^2 + 4 \|\gamma_2 \nabla_x V_{\theta_1}(x) - \gamma_1 \nabla_x V_{\theta_1}(x)\|^2 \\ &\qquad + 4 \left\|\gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x)\right\|^2 + 4 \left\|\gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x)\right\|^2 \; . \end{aligned} \tag{47}$$

First using **H**4 we have

$$\|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_2 \nabla_x V_{\theta_1}(x)\| \leqslant \gamma_2 \mathtt{M}_\Theta \|\theta_1 - \theta_2\| (1 + \|x\|) \; . \tag{48}$$

Second using **H**1 we have

$$\begin{aligned} \|\gamma_2 \nabla_x V_{\theta_1}(x) - \gamma_1 \nabla_x V_{\theta_1}(x)\| &\leqslant (\gamma_1 - \gamma_2) \|\nabla_x V_{\theta_1}(x)\| \\ &\leqslant (\gamma_1 - \gamma_2) \mathtt{L} \left\|x - x_{\theta_1}^\star\right\| \leqslant (\gamma_1 - \gamma_2) \mathtt{L}(R_{V,1} + \|x\|) \; . \end{aligned} \tag{49}$$

Third using **H**1, **H**4, Lemma 5 and Lemma 7 we have

$$\begin{aligned} \left\|\gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x)\right\| &\leqslant \left\|(x - \mathrm{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x))/\kappa - (x - \mathrm{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x))/\kappa\right\| \\ &\leqslant \left\|\mathrm{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x) - \mathrm{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x)\right\| \Big/ \kappa \\ &\leqslant 2\mathtt{M}(\gamma_1 - \gamma_2) \; . \end{aligned} \tag{50}$$

Finally using **H**4 we have

$$\left\|\gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x)\right\| \leqslant \gamma_2 \left\{ \sup_{[0,\bar{\gamma}\kappa]} \mathtt{f}_\theta(t) \right\} \|\theta_1 - \theta_2\| \; . \tag{51}$$

Combining (48), (49), (50) and (51) in (47) we get that there exists $\bar{A}_{4,1} \geqslant 0$ such that

$$\begin{aligned} &\|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_1 \nabla_x V_{\theta_1}(x) + \gamma_2 \nabla_x U_{\theta_2}^{\kappa}(x) - \gamma_1 \nabla_x U_{\theta_1}^{\kappa}(x)\|^2 \\ &\qquad \leqslant \bar{A}_{4,1} \left[(\gamma_1 - \gamma_2)^2 + \gamma_2^2 \|\theta_1 - \theta_2\|\right] W^{2a}(x) \; . \end{aligned}$$

Using this result in (46) we obtain that there exists $\bar{A}_{4,2} \geqslant 0$ such that

$$\mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right) \leqslant \bar{A}_{4,2} \left[(\gamma_1/\gamma_2 - 1) + \gamma_2 \|\theta_1 - \theta_2\|^2\right] W^{2a}(x) \; ,$$

which implies the announced result upon setting $\bar{A}_4 = 2\sqrt{\bar{A}_{4,2}}(1 + b\bar{\gamma})^{1/2}$ and using that for any $u, v \geqslant 0$, $\sqrt{u+v} \leqslant \sqrt{u} + \sqrt{v}$. □

D.4 Proof of Theorem 1

We divide the proof in two parts.

(a) First assume that $(X_k^n)_{n \in \mathbb{N}, k \in \{0,\ldots,m_n\}}$ and $(\bar{X}_k^n)_{n \in \mathbb{N}, k \in \{0,\ldots,m_n\}}$ are given by (23) with $\{(\mathrm{K}_{\gamma,\theta}, \bar{\mathrm{K}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(\mathrm{S}_{\gamma,\theta}, \bar{\mathrm{S}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$. Then Lemma 15 implies that [80, H1a] is satisfied with $A_1 \leftarrow A_1$, Theorem 17 implies that [80, H1b] holds with $A_2 \leftarrow A_2$ and $\rho \leftarrow \rho$. Finally, using

Corollary 22 we get that [80, H1c] holds with $\boldsymbol{\Psi} \leftarrow \boldsymbol{\Psi}$. Therefore, we can apply [80, Theorem 1] and we obtain that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. if

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

Since $\boldsymbol{\Psi}(\gamma_n) = \mathcal{O}(\gamma_n^{1/2})$ by Corollary 22, these summability conditions are satisfied under the summability assumptions of Theorem 1 (1). Proposition 23 implies that [80, H2] holds with $\boldsymbol{\Lambda}_1 \leftarrow \boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2 \leftarrow \boldsymbol{\Lambda}_2$. Therefore if $m_n = m_0$ for all $n \in \mathbb{N}$, we can apply [80, Theorem 3] and we obtain that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. if

$$\sum_{n=0}^{+\infty} \delta_n = +\infty \,, \qquad \sum_{n=0}^{+\infty} \delta_{n+1} \boldsymbol{\Psi}(\gamma_n) < +\infty \,, \qquad \sum_{n=0}^{+\infty} \delta_{n+1} \gamma_n^{-2} < +\infty$$
$$\sum_{n=0}^{+\infty} \delta_{n+1}/\gamma_n^2 (\boldsymbol{\Lambda}_1(\gamma_n, \gamma_{n+1}) + \delta_{n+1} \boldsymbol{\Lambda}_2(\gamma_n, \gamma_{n+1})) < +\infty \,.$$

These summability conditions are satisfied under the summability assumptions of Theorem 1 (2).

(b) Second assume that $(X_k^n)_{n\in\mathbb{N},k\in\{0,\dots,m_n\}}$ and $(\bar{X}_k^n)_{n\in\mathbb{N},k\in\{0,\dots,m_n\}}$ are given by (23) with $\{(\mathrm{K}_{\gamma,\theta}, \bar{\mathrm{K}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}] \,, \theta \in \Theta\} = \{(\mathrm{R}_{\gamma,\theta}, \bar{\mathrm{R}}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}] \,, \theta \in \Theta\}$. Then Lemma 26 implies that [80, H1a] is satisfied with $A_1 \leftarrow \bar{A}_1$, Theorem 28 implies that [80, H1b] holds with $A_2 \leftarrow \bar{A}_2$ and $\rho \leftarrow \bar{\rho}$. Finally, using Proposition 31 we get that [80, H1c] holds with $\boldsymbol{\Psi} \leftarrow \bar{\boldsymbol{\Psi}}$. Therefore, we can apply [80, Theorem 1] and we obtain that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. if

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

Since $\boldsymbol{\Psi}(\gamma_n) = \mathcal{O}(\gamma_n^{1/2})$ by Proposition 31, these summability conditions are satisfied under the summability assumptions of Theorem 1 (1). Proposition 32 implies that [80, H2] holds with $\boldsymbol{\Lambda}_1 \leftarrow \bar{\boldsymbol{\Lambda}}_1$ and $\boldsymbol{\Lambda}_2 \leftarrow \bar{\boldsymbol{\Lambda}}_2$. Therefore if $m_n = m_0$ for all $n \in \mathbb{N}$, we can apply [80, Theorem 3] and we obtain that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. if

$$\sum_{n=0}^{+\infty} \delta_n = +\infty \,, \quad \sum_{n=0}^{+\infty} \delta_{n+1} \bar{\boldsymbol{\Psi}}(\gamma_n) < +\infty \,, \quad \sum_{n=0}^{+\infty} \delta_{n+1}^2 \gamma_n^{-2} \,,$$
$$\sum_{n=0}^{+\infty} \delta_{n+1}/\gamma_n^2 (\bar{\boldsymbol{\Lambda}}_1(\gamma_n, \gamma_{n+1}) + \delta_{n+1} \bar{\boldsymbol{\Lambda}}_2(\gamma_n, \gamma_{n+1})) < +\infty \,.$$

These summability conditions are satisfied under the summability assumptions of Theorem 1 (2).

D.5 Proof of Theorem 2

The proof is similar to the one of Theorem 1 using [81, Theorem 2, Theorem 4] instead of [81, Theorem 1, Theorem 3].