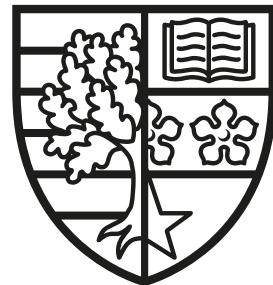


# **Bayesian computation in imaging inverse problems with partially unknown models**

Ana Fernandez Vidal

SUBMITTED FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

HERIOT-WATT UNIVERSITY



DEPARTMENT OF MATHEMATICS,  
SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES.

November 2020

The copyright in this thesis is owned by the author. Any quotation from the thesis or use of any of the information contained in it must acknowledge this thesis as the source of the quotation or information.

## Abstract

Many imaging problems require solving a high-dimensional 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 thesis, 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, and demonstrate it 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  $\ell_1$ , total-variation, total-variation and  $\ell_1$ , and total-generalised-variation pseudo-norms. Moreover, we explore some other applications of the proposed method including maximum marginal likelihood estimation in Bayesian logistic regression and audio compressed sensing, as well as an application to model selection based on residuals.

## Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Dr. Marcelo Pereyra, for his great guidance, encouragement, patience and support throughout my Ph.D. studies. He has always found the time, however busy, to discuss my work, reply to my emails (almost scarily fast), and share his expertise.

Besides my supervisor, I would like to extend my sincere gratitude to my external examiner, Prof. Josiane Zerubia, and my internal examiner, Prof. Yoann Altmann, for reviewing my work with such care and dedication. Both had extremely interesting and insightful questions and observations that helped me improve this manuscript and triggered many new ideas for future research.

I would also like to extend my sincere thanks to my second supervisor Prof. Gavin Gibson and Prof. Yoann Altmann, for their insightful comments and support throughout these three years.

I also had the great pleasure of working with Valentin De Bortoli and Dr. Alain Durmus. I really appreciate the effort they have made to coordinate our collaboration remotely from France.

I would like to thank my friends at Heriot-Watt for their moral support and for all the nice moments we shared in the last three years.

I especially want to thank my friend Rui, who has not only been an outstanding friend but has also spent countless hours proof-reading my articles, giving me feedback and helping me implement my work in Julia.

Last but not least, I would like to thank my husband Robin and my cat John for their constant love and support throughout writing this thesis and in my life in general.

## Research Thesis Submission

Name:	Ana Fernandez Vidal		
School:	Mathematical and Computer Sciences		
Version: (i.e. First, Resubmission, Final)	Final	Degree Sought:	Ph.D. in Statistics

### Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1. The thesis embodies the results of my own work and has been composed by myself
2. Where appropriate, I have made acknowledgement of the work of others
3. The thesis is the correct version for submission and is the same version as any electronic versions submitted\*.
4. My thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5. I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.
6. I confirm that the thesis has been verified against plagiarism via an approved plagiarism detection application e.g. Turnitin.

### ONLY for submissions including published works

7. Where the thesis contains published outputs under Regulation 6 (9.1.2) or Regulation 43 (9) these are accompanied by a critical review which accurately describes my contribution to the research and, for multi-author outputs, a signed declaration indicating the contribution of each author (complete)
8. Inclusion of published outputs under Regulation 6 (9.1.2) or Regulation 43 (9) shall not constitute plagiarism.

\* Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.

Signature of Candidate:		Date:	17-12-2020
-------------------------	---	-------	------------

### Submission

Submitted By (name in capitals):	Ana Fernandez Vidal		
Signature of Individual Submitting:			
Date Submitted:	15-01-2021		

### For Completion in the Student Service Centre (SSC)

Limited Access	Requested	Yes	No	Approved	Yes	No
E-thesis Submitted (mandatory for final theses)						
Received in the SSC by (name in capitals):				Date:		

## Inclusion of Published Works

---

### Declaration

This thesis contains one or more multi-author published works. In accordance with Regulation 6 (9.1.2) I hereby declare that the contributions of each author to these publications is as follows:

Citation details	A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part I: Methodology and Experiments, SIAM Journal on Imaging Sciences, 13(4), 1945-1989 (2020).
A. F. Vidal	Designed, implemented and evaluated the proposed methodology. Carried out all the experiments in MATLAB. Drafting, writing and proof-reading of the article.
V. De Bortoli	Contributed to reviewing the technical parts of the paper, and defined all the relevant notation.
M. Pereyra	Conception of the method. Supervision and guidance in all stages. Writing and revision of the article.
A. Durmus	Reviewed all technical aspects of the article.
Signature:	
Date:	17-12-2020

Citation details	V. De Bortoli, D. Alain, M. Pereyra, and A. F. Vidal, Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part II: Theoretical Analysis, SIAM Journal on Imaging Sciences, 13(4), 1990-2028 (2020).
V. De Bortoli	Theoretical analysis of the proposed methodology, drafting of full article, writing of all the technical sections of the article.
A. Durmus	Theoretical analysis of the proposed methodology, revision of all the technical sections of the article and proofs. General revision and writing of full article.
M. Pereyra	Contributed to writing all the non-technical parts of the paper, especially introduction. Revision of some of the technical sections.
A. F. Vidal	Contributed to drafting some non-technical parts of the paper, especially introduction. Proof-reading.
Signature:	
Date:	17-12-2020

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Inverse Problems in Imaging Applications . . . . .	1
1.2	Regularisation in a Bayesian Framework . . . . .	3
1.3	Setting regularisation parameters . . . . .	6
1.4	Contributions . . . . .	8
1.5	Outline . . . . .	10
1.6	Connections to other approaches and frameworks . . . . .	11
1.7	Limitations of the existing approaches . . . . .	15
1.8	Publications . . . . .	17
1.9	Other research activities . . . . .	18
<b>2</b>	<b>Overview of existing methods for selecting regularisation parameters</b>	<b>19</b>
2.1	Methods based on residual analysis . . . . .	19
2.2	Methods based on surrogates of the MSE . . . . .	22
2.3	Bayesian methods . . . . .	23
2.3.1	Hierarchical Bayesian estimation . . . . .	23
2.3.2	Empirical Bayes estimation . . . . .	26
2.3.3	Connections between both Bayesian approaches . . . . .	27
<b>3</b>	<b>Proposed methodology</b>	<b>29</b>
3.1	Proposed algorithm . . . . .	30
3.1.1	Scalar-valued $\theta$ with homogeneous regulariser . . . . .	32
3.1.2	Separably homogeneous regulariser . . . . .	33
3.1.3	General case: inhomogeneous regulariser . . . . .	35

3.1.4	MCMC Kernels . . . . .	35
3.1.5	Connections to the expectation-maximisation algorithm . . . . .	39
3.2	Example on synthetic data . . . . .	41
3.2.1	Estimation variance and bias . . . . .	41
3.2.2	Laplace noise and likelihood misspecification . . . . .	43
3.2.3	Role of the algorithm parameters . . . . .	45
3.3	Implementation guidelines . . . . .	47
3.3.1	Setting the algorithm parameters . . . . .	48
3.3.2	Other implementation considerations . . . . .	51
3.3.3	Testing the MCMC sampler . . . . .	51
3.3.4	Monitoring convergence in Algorithm 1, Algorithm 2 and Algorithm 3 . . . . .	52
3.3.5	Working with two MCMC chains in Algorithm 3 . . . . .	53
3.3.6	Working with multivariate $\theta$ . . . . .	54
3.3.7	Convergence speed . . . . .	54
3.3.8	Estimation Bias . . . . .	55
<b>4</b>	<b>Numerical experiments on imaging problems</b>	<b>57</b>
4.1	Non-blind natural image deconvolution . . . . .	58
4.1.1	Deconvolution with total variation prior . . . . .	59
4.1.2	Deconvolution with TV prior and unknown noise variance . . . . .	62
4.1.3	Wavelet deconvolution with synthesis prior . . . . .	66
4.2	Hyperspectral Unmixing with TV-SUnSAL . . . . .	69
4.3	Denoising with a total generalised variation prior . . . . .	74
<b>5</b>	<b>Beyond imaging applications</b>	<b>83</b>
5.1	Generalised SAPG algorithm . . . . .	83
5.2	Bayesian Logistic Regression . . . . .	85
5.3	Audio compressed sensing . . . . .	89
5.4	Bayesian logistic regression with random effects . . . . .	91
<b>6</b>	<b>Model selection</b>	<b>95</b>
6.1	Introduction . . . . .	95
6.2	Bayesian model selection . . . . .	96

*CONTENTS*

6.3	Proposed model selection method . . . . .	98
6.4	Numerical experiments . . . . .	98
6.5	Conclusions . . . . .	103
<b>7</b>	<b>Conclusions and perspectives for future work</b>	<b>105</b>
<b>A</b>	<b>Fisher's identity</b>	<b>108</b>
<b>B</b>	<b>Fair comparison of different methods for setting <math>\theta</math></b>	<b>109</b>
B.1	Comparing with solver-dependent methods . . . . .	110
<b>C</b>	<b>Analysis of the convergence properties</b>	<b>112</b>
	<b>Bibliography</b>	<b>143</b>

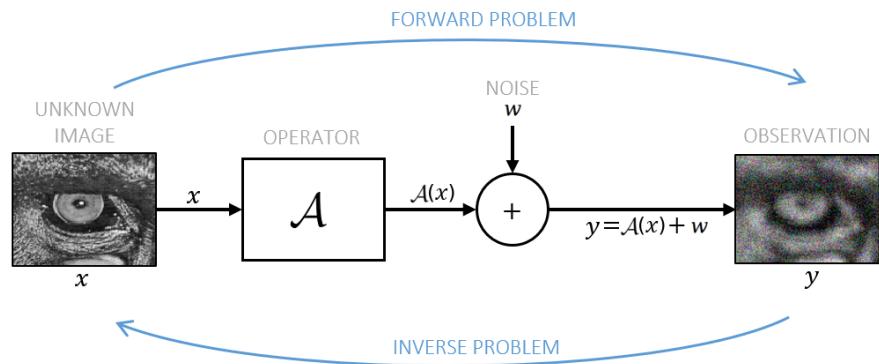
# Chapter 1

## Introduction

### 1.1 Inverse Problems in Imaging Applications

Mathematical imaging is at the core of modern data science, with important applications in medicine, biology, defence, agriculture and environmental sciences. Many of the problems arising in these disciplines involve the estimation of an unobserved image  $x \in \mathbb{C}^d$ , from measurements  $y \in \mathbb{C}^{d_y}$  that are noisy, incomplete and resolution-limited. Canonical examples include, for instance, image denoising [89], image de-blurring [39, 80], compressed sensing [53, 115], super-resolution [100, 145], tomographic reconstruction [40], image inpainting [62, 125], source separation [15, 44], fusion [90, 126], and phase retrieval [26, 76].

For example, in many applications the observed vector  $y$  can be modelled as the output of a system  $\mathcal{A}$  such that  $y = \mathcal{A}(x) + w$ , where  $\mathcal{A}$  is an operator that describes how the underlying image  $x$  gives rise to  $y$ , and  $w$  is some observation noise.



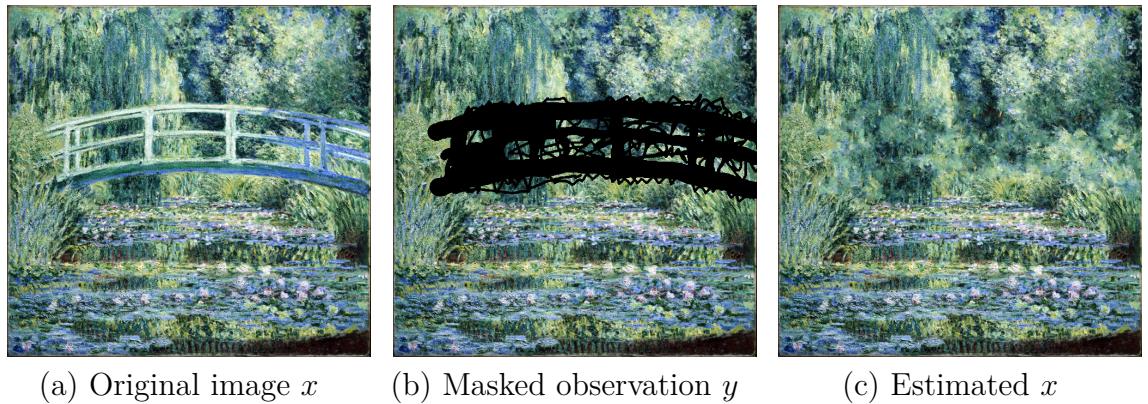
**Figure 1.1** – In a forward problem, the transformation  $\mathcal{A}$  is applied to an input image  $x$ . The inverse problem aims to obtain an estimate of  $x$  from the noisy observation  $y$ .

The operator  $\mathcal{A}$  can be expressed in different ways (differential equations, integral equations or other mathematical mappings) and it can represent transformations such as the blur introduced by the lens of an imaging device, a random sampling operator or an inpainting degradation mask modelling lost pixels [92].

Given an input  $x$  and an operator  $\mathcal{A}$ , determining the output  $y$  represents the **forward problem**. Finding the input  $x$  for a given output  $y$  represents the **inverse problem** (see Figure 1.1).

Solving such inverse problems can be a substantially difficult task. The properties of the mapping  $\mathcal{A}$  can make the recovery of  $x$  from  $y$  very challenging, often leading to problems that are **ill-posed** or **ill-conditioned**. A problem is *ill-posed* if it is not well-posed in the sense of Hadamard [67], i.e. if there is no unique solution that continuously depends on the observed data  $y$ . Even when the problem is well-posed, it can still be *ill-conditioned* if small perturbations in  $y$  result in large perturbations in the estimated  $x$ , making the problem very sensitive to noise.

In many imaging problems  $\mathcal{A}$  has a non-trivial nullspace, either because it is rank-deficient or because  $d_y$  is smaller than  $d$ , leading to ill-posed problems that do not have a unique solution. For example, consider the inpainting problem in Figure 1.2, where the observed image  $y$  has some missing or occluded parts, and the goal is to use the information in  $y$  to estimate the values of the unobserved pixels of  $x$ . Since the unobserved pixels do not directly affect the observed pixels, there is an infinite number of possible images  $x$  that are compatible with a given observation  $y$ .



**Figure 1.2** – Inpainting of Monet’s Water Lilies from a masked observation (b), using a Mathematica implementation of the texture synthesis method proposed in [70].

Likewise, the spectral properties of the mapping  $\mathcal{A}$  can sometimes make the re-

covery of  $x$  very unstable: when the quotient between the maximal and the minimal singular values of  $\mathcal{A}$  is too large, the problem is ill-conditioned and even mild observational noise can lead to extremely noisy estimates. This is often the case with, for instance, deblurring problems as the one illustrated in Figure 1.3 (unstable solution shown on the left).

In spite of all these difficulties, joint research efforts over the last decades have led to impressive advances in methods, models and algorithms for imaging inverse problems [92, 99, 108, 125].

## 1.2 Regularisation in a Bayesian Framework

There are many different mathematical frameworks available to address such imaging problems [4, 78, 92]. Despite the broad range of models and applications, most imaging methods adopt a similar strategy for approaching these ill-posed/ill-conditioned problems: they render them well-posed by augmenting the information in the observation using additional knowledge about the signal to be recovered. This process is called **regularisation** and it can be attained in many different ways.

In this thesis, we adopt a **Bayesian statistical framework** where regularisation arises from the use of informative prior distributions that promote solutions with expected structural or regularity properties (e.g., smoothness, piecewise-regularity, sparsity, etc.). More precisely, we focus on problems where the observation  $y$  is related to  $x$  by a statistical model with likelihood function

$$p(y|x) \propto e^{-f_y(x)}, \quad (1.1)$$

where  $f_y$  is convex and continuously differentiable with  $L_y$ -Lipschitz gradient, *i.e.* for any  $u, v \in \mathbb{R}^d$ ,  $\|\nabla f_y(u) - \nabla f_y(v)\| \leq L_y \|u - v\|$  where  $L_y > 0$ . This class includes important observation models, in particular Gaussian linear models of the form  $y = Ax + w$  where  $A \in \mathbb{C}^{d_y \times d_x}$ ,  $w \sim \mathcal{N}(0, \sigma^2 I_{d_y})$  with  $\sigma > 0$ , and  $f_y(x) = \frac{\|y - Ax\|_2^2}{2\sigma^2}$ . We want to stress at this point that this work assumes that  $f_y(x)$  is known, which implies that the noise variance  $\sigma^2$  is also known.

Let  $\Theta \subset (0, +\infty)^{d_\theta}$  be a convex compact set. Following a Bayesian approach, we model our prior knowledge about  $x$  using a prior distribution given for any  $\theta \in \Theta$

by

$$p(x|\theta) = \frac{e^{-\sum_{i=1}^{d_\Theta} \theta_i g_i(x)}}{Z(\theta)} = \frac{e^{-\theta^T g(x)}}{Z(\theta)}, \quad (1.2)$$

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

$$Z(\theta) = \int_{\mathbb{R}^d} e^{-\theta^T g(\tilde{x})} d\tilde{x}. \quad (1.3)$$

Note that  $\theta$  is a multivariate regularisation parameter that controls the amount of regularity enforced. The function  $g$  is allowed to be non-differentiable in order to include popular models such as  $g(x) = \|Bx\|_1$  for some dictionary  $B \in \mathbb{R}^{d_1 \times d}$  with  $d_1 \in \mathbb{N}$  and norm  $\|\cdot\|_1$ , as well as constraints on the solution space such as pixel-positivity. One could also consider more complex priors, such as plug-and-play priors defined via denoising algorithms [135], but then the theoretical guarantees that we establish in Appendix C might not hold.

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  $\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|\theta) d\tilde{x}$ , which depends on  $\theta$ . Formally, by differentiating (1.3) and using Leibniz integral rule [107] we obtain that for any  $\theta \in \Theta$

$$\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|\theta) d\tilde{x} = -\nabla_\theta \log Z(\theta). \quad (1.4)$$

Additionally, because the prior distribution  $x \mapsto p(x|\theta)$  is log-concave, using [16, Theorem 1.2] we have that for any  $\varepsilon \in [0, 2]$

$$\int_{C_{\theta,\varepsilon}} p(\tilde{x}|\theta) d\tilde{x} \leq 3 \exp[-\varepsilon^2 d / 16], \quad (1.5)$$

with  $C_{\theta,\varepsilon} = \{\tilde{x} \in \mathbb{R}^d : d^{-1} |\theta^T(g(\tilde{x}) - \bar{g}_\theta)| \geq \varepsilon\}$ . This result establishes that the prior distribution  $x \mapsto p(x|\theta)$  strongly concentrates the probability mass on solutions for which  $g(x) \approx -\nabla_\theta \log Z(\theta)$  with high probability when  $d$  is large. In other words, when the dimension of  $x$  is high, the prior distribution promotes values of  $x$  for which  $g(x)$  is very close to its expectation  $\bar{g}_\theta$ , and the value of  $\bar{g}_\theta$  is directly determined by  $\theta$  through (1.4).

Once the likelihood and prior  $p(y|x)$  and  $p(x|\theta)$  are specified, we use Bayes'

theorem [118] to derive the posterior for any  $\theta \in \Theta$  and  $x \in \mathbb{R}^d$

$$p(x|y, \theta) = p(y|x)p(x|\theta)/p(y|\theta) = \exp[-f_y(x) - \theta^T g(x)] / \int_{\mathbb{R}^d} \exp[-f_y(\tilde{x}) - \theta^T g(\tilde{x})] d\tilde{x} . \quad (1.6)$$

This posterior distribution underpins all inferences about the image  $x$  given observed data  $y$ , and it can be used in different ways to obtain estimates of  $x$ . In particular, imaging methods often use the maximum-a-posteriori (MAP) estimator, given for any  $\theta \in \Theta$  by

$$\hat{x}_{\theta, \text{MAP}} \in \operatorname{argmin}_{\tilde{x} \in \mathbb{R}^d} \{f_y(\tilde{x}) + \theta^T g(\tilde{x})\} . \quad (1.7)$$

This Bayesian estimator has a number of favourable theoretical and computational properties (see [104] 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  $\hat{x}_{\theta, \text{MAP}}$  is a convex optimisation problem that can usually be efficiently solved using modern optimisation algorithms, see [36]. Imaging MAP algorithms typically adopt a proximal splitting approach [42] involving the gradient  $\nabla f_y$  and the proximal operator of  $g$ ,  $\operatorname{prox}_g^\lambda : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\Theta}$ , see [13, Definition 12.23]. This operator is defined for any  $\lambda > 0$  and  $x \in \mathbb{R}^d$  by

$$\operatorname{prox}_g^\lambda(x) = \operatorname{argmin}_{\tilde{x} \in \mathbb{R}^d} \{g(\tilde{x}) + \|\tilde{x} - x\|_2^2/(2\lambda)\} , \quad (1.8)$$

The smoothness parameter  $\lambda > 0$  controls the regularity properties of the proximal operator.

It is worth noting that the posterior (1.6) can also be used to compute other estimators of  $x$ . For example, one can compute the maximiser of the posterior marginals (MPM) [95] or perform minimum mean squared error (MMSE) estimation [118] by computing the posterior mean

$$\hat{x}_{\text{MMSE}} = \int_{\mathbb{R}^d} x p(x|y, \theta) dx. \quad (1.9)$$

Unlike MAP estimation, which is predominantly computed by using optimisation algorithms, the MMSE estimator and other Bayesian estimators are generally not available as optimisation problems and need to be calculated with other methods. There is a wide range of alternative computational techniques, see [63, 66, 108].

As mentioned previously, the regularisation parameter  $\theta \in \Theta$  controls the region where the prior probability mass is concentrated, and this can significantly impact inferences about the unknown image  $x \in \mathbb{R}^d$ , especially in problems that are ill-posed or ill-conditioned.

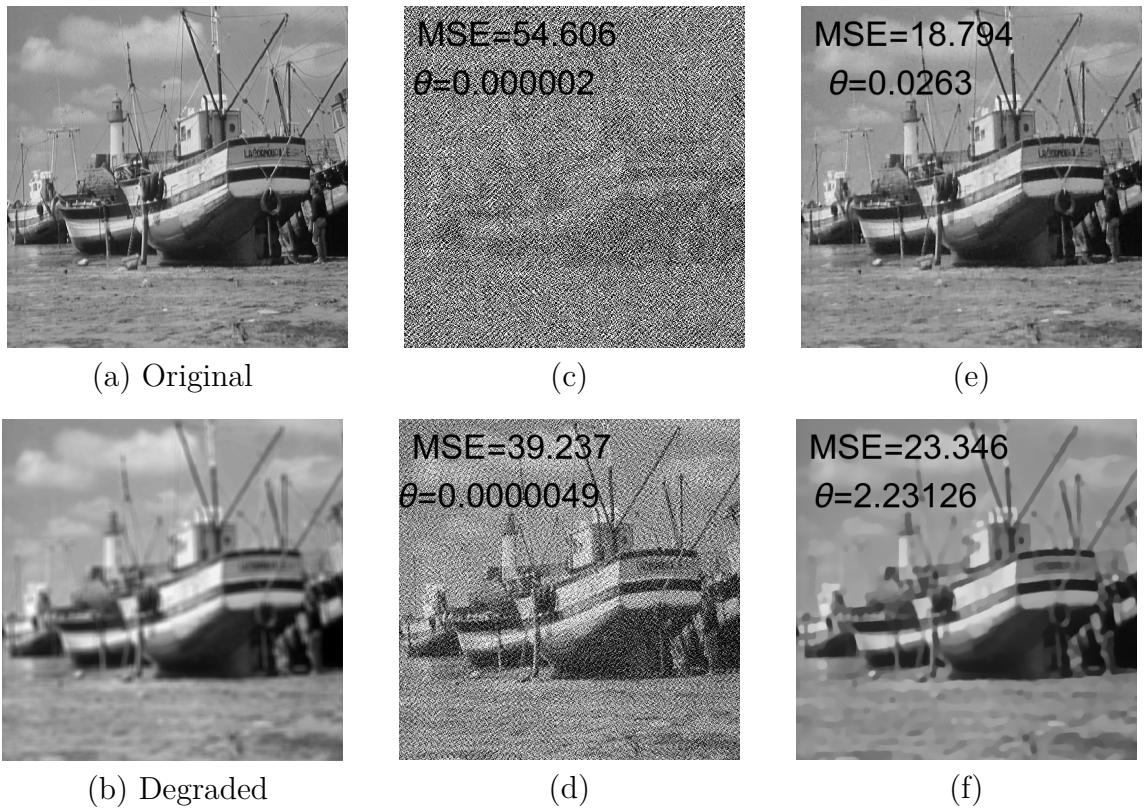
## 1.3 Setting regularisation parameters

A main difficulty that arises when using most regularisation techniques is deciding how much regularisation is appropriate. Different imaging modalities, instrumental setups, scenes, and noise conditions often require using very different amounts of regularisation. As previously explained, the amount of regularisation is usually controlled by the regularisation parameter  $\theta$ , and setting its value can be very difficult (see [49, 98, 105]).

In Figure 1.3, 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  $\theta$  is too small the estimated image is very noisy due to lack of regularisation, and when  $\theta$  is too large the resulting image is over-regularised.

We want to mention at this point that this difficulty is not inherent to Bayesian approaches. Other regularisation techniques used with different mathematical frameworks face the same challenge when it comes to setting regularisation parameters (in Section 1.6 we introduce some of these alternative approaches and discuss connections to the Bayesian framework). As a result, there is significant interest in methods for setting regularisation parameters in an automatic, robust, and adaptive way, and this will be the main focus of this thesis.

Indeed, the developments of methods to automatically set regularisation parameters is a long-standing research topic in imaging sciences. Some methods such as generalised cross-validation [65], the L-curve [23, 69, 88], the discrepancy principle [14, 101] and residual whiteness measures [3, 87] 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



**Figure 1.3** – Deblurring of the **boat** image with total-variation prior: (a) True image  $x$ , (b) blurred ( $9 \times 9$  uniform blur) and noisy observation  $y$  (SNR=40 dB), and (c-f) maximum-a-posteriori estimators for different values of  $\theta > 0$  illustrating the effect of regularisation (increasing from (c) to (f)).

of attention recently [49, 57, 64]. These methods seek to select the value of the regularisation parameters by minimising SURE-based surrogates of the estimation mean squared error [57, 64, 110]. 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 [93]. Some recent works also consider learning regularisation parameters from a training dataset of clean images [131], or adopting a bilevel optimisation strategy [22, 83].

Lastly, the Bayesian statistical framework provides two main strategies for addressing unknown regularisation parameters: the hierarchical and the empirical [98, 118]. 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 [105, 108]. This is the strategy that is adopted by most Markov chain Monte Carlo and variational

Bayesian approaches reported in the literature (see e.g., [10, 106]). In contrast, the empirical Bayesian approach has been studied relatively little because it involves solving an intractable maximum marginal likelihood estimation (MMLE) problem (see Section 2.3.2).

In this thesis 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. The main contributions are summarised below.

## 1.4 Contributions

The main contributions of this thesis are:

1. **A new method for setting regularisation parameters:** we propose an empirical Bayesian method to estimate the regularisation parameters directly from the observed data by maximum marginal likelihood estimation. A main novelty is that this maximum marginal likelihood estimation problem is 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.  
The algorithm is very general, computationally efficient and easy to implement (it only requires knowing gradient and proximal operators so it is straightforward to apply to problems that are currently solved by proximal optimisation). Moreover, it can be used to estimate multiple parameters simultaneously (most alternative approaches from the literature are for scalar parameters). We propose two main versions of the methodology: one for problems with tractable partition functions or homogeneous regularisers, which requires a single Markov kernel targeting the posterior distribution of  $x$ , and one for all other cases, which employs two different Markov kernels targeting both the posterior and the prior distributions of  $x$  and thus requires the prior to be proper.
2. **Detailed practical guidelines:** bridging the gap between theory and practice has been one of the core goals of this thesis. Although similar ideas to the ones proposed in this thesis have been studied in recent works [6, 60], they

have mostly focused on the theory rather than the practical aspects. Hence it is very difficult for practitioners to, for example, check if the random samplers they use verify the necessary conditions for convergence. Moreover, in some cases those methods require incrementing the number of random samples used at each iteration, which is not always feasible in practice (our proposed method works with only one random sample per iteration). One of the most important contributions of this work is that the proposed method has been designed to be practical from the practitioner’s point of view, and that we provide very comprehensive implementation guidelines that cover every aspect needed to get this method running (how to test the sampler, how to set up every parameter, how to troubleshoot, etc.).

3. **Theoretical analysis:** we 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 work of this thesis has been carried out in close collaboration with applied probability experts Valentin De Bortoli and Alain Durmus, who had a leading role in this theoretical analysis.
4. **Numerical experiments:** we demonstrate the proposed methodology with a broad range of severely ill-posed imaging problems as well as some other statistical problems such as logistic regression or audio compressive sensing. The method is very robust to noise and delivers remarkably accurate solutions, usually outperforming other alternative approaches.
5. **Noise variance estimation:** As mentioned earlier, throughout this work we assume that the noise variance  $\sigma^2$  is known. This is a standard assumption in the literature (see, e.g., [49, 105]) that is sometimes difficult to verify in practice. To mitigate this issue, in Section 4.1.2 we study a possible way to incorporate the estimation of  $\sigma^2$  into the proposed scheme.
6. **Model selection:** using the proposed empirical Bayesian method, we introduce a fast heuristic for comparing Bayesian models to solve inverse problems where no ground truth is available. The proposed heuristic is very computationally efficient and does not require the estimation of the model evidence. We illustrate this approach for model selection with a total-variation image

deblurring experiment, where it performs remarkably well.

## 1.5 Outline

The thesis is organised as follows:

In the remainder of Chapter **1** we discuss the context of our contributions. We briefly introduce other approaches and frameworks for solving ill-posed inverse problems and we make connections to the contributions of this work. In particular, we explain why the proposed methodology can also be used for setting regularisation parameters in other non-Bayesian approaches and we end by reviewing current perspectives and limitations of the existing approaches to solving inverse problems.

Chapter **2** provides an overview of the available methods for selecting the regularisation parameters and discusses the connections between empirical and hierarchical Bayesian approaches.

Chapter **3** presents the proposed empirical Bayesian method to calibrate regularisation parameters. The algorithm is presented in three different versions depending on the properties of the regulariser and the tractability of the partition function. We use a synthetic image denoising problem to study the behaviour of the algorithm in depth and we provide detailed implementation and troubleshooting guidelines. We also discuss connections to the expectation-maximisation algorithm.

In Chapter **4** we demonstrate the proposed methodology on a broad range of ill-posed and ill-conditioned imaging inverse problems. We first consider different non-blind image deblurring problems involving scalar-valued regularisation parameters, including an experiment where we also estimate the noise variance. This is followed by a challenging sparse hyperspectral image unmixing with the SUnSAL model [75], which involves two different regularisation parameters. Finally, we consider a more challenging denoising problem with a Total Generalised Variation regulariser [18] that requires setting vector-valued regularisation parameters that have strong dependencies, making the estimation problem particularly difficult. We report comparisons with several alternative approaches from the literature, including the discrepancy principle [101], the SURE-based SUGAR method [49], and the hierarchical Bayesian method from [105].

In Chapter 5 we show the scope of the proposed methodology for estimating other kinds of parameters in other types of non-imaging inverse problems. We first present a generalised version of the proposed method that can be applied to a broader range of intractable maximum likelihood estimation problems and then illustrate its performance with an audio compressive sensing problem and with two Bayesian logistic regression problems with and without random effects.

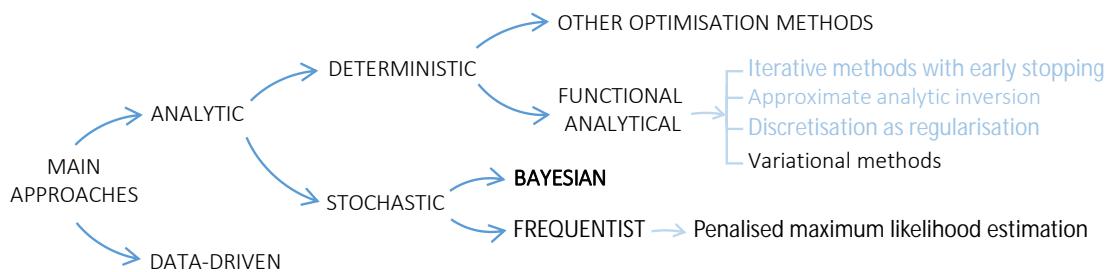
In Chapter 6, we introduce a different application of the proposed methodology: we propose a fast heuristic for comparing Bayesian models under no ground truth. The proposed model selection method is illustrated with a total-variation image deblurring experiment, where it performs remarkably well.

Conclusions and perspectives for future work are finally reported in Chapter 7.

Appendix C presents a detailed analysis of the theoretical properties of the proposed methodology, including easily verifiable conditions for convergence and quantitative convergence rates.

## 1.6 Connections to other approaches and frame-works

We now provide a brief overview of the main approaches to address ill-posed imaging problems and discuss how the proposed methodology could be used for setting regularisation parameters in some of these non-Bayesian approaches. Figure 1.4



**Figure 1.4** – Different approaches to solving inverse problems.

sketches out the main families of approaches. One of the defining features at the core of an approach is the way in which the forward problem is modelled. While **analytical techniques** define an explicit forward model, for instance by defining an operator  $\mathcal{A}$ , **data-driven techniques** [92] can learn the relations between the

underlying image and the observations without the need to define a detailed forward model. In this case, the structure is learnt from data rather than specified analytically. The most common way of attaining this, is by using (deep) neural networks to minimise some cost functional  $\|x - g_\Phi(y)\|$ , where  $g_\Phi$  tries to approximately invert the forward model and the noise, and  $\Phi$  represents the set of neural network parameters. These parameters are learnt from data by training the neural network with the available ground truth (i.e. a large set of  $(x, y)$  pairs).

Both analytical and data-driven approaches have some inherent limitations. Data-driven techniques do not offer a straight-forward way of incorporating such prior knowledge and tend to be limited by what can be learnt from the data. For instance, if a parameter in the system of interest changes (e.g. a microscope lens), data-driven approaches typically have to re-learn the whole model instead of tuning a specific model parameter. On the other hand, it is often hard to obtain realistic regularisers or prior distributions that can be written down as a simple mathematical expression. In most high-dimensional problems the unknown signals are typically concentrated on lower dimensional subspaces of  $\mathbb{R}^d$ . Capturing the exact structure of these subspaces is a very challenging problem which often exceeds the capabilities of analytic methods.

To overcome these limitations, recent works [4] have started to explore the fusion between analytic and data-driven models. In particular, a popular approach is the one of Plug-and-Play priors [122, 135], where a learnt denoiser is used instead of a proximal operator in optimisation schemes such as the ones used to compute (1.7). In this case one could try to estimate the parameters of such PnP priors using the proposed methodology. We have not tested this in this thesis, but it is part of the perspectives for future work. We expect that our method would work in some cases but that it will be very difficult to develop theoretical guarantees with these learnt priors.

Within the analytic models, **deterministic approaches** usually formulate the inverse recovery as an optimisation problem which they regularise by adding a penalty term to the target cost to promote solutions with desired properties. In particular, **variational approaches**<sup>1</sup> follow this strategy in a continuous space:

---

<sup>1</sup>Variational approaches are just one of four main regularisation strategies within the functional analytic framework. For more details see [4].

adopting a functional analytic framework, they model images as functions rather than a finite-dimensional vector of pixels, and then compute point estimates for the unknown image as minimisers of a cost functional with both data-fidelity and regularisation terms, that is,

$$\hat{x}_f = \operatorname{argmin}_{x_f \in \mathcal{X}} V_f(\mathcal{A}_f(x_f), y_f) + \theta^T g_f(x_f) \quad (1.10)$$

where  $x_f$  is a continuous image function defined in a Banach space  $\mathcal{X}$ , and  $V_f$  is a cost function that plays a similar role to the likelihood in a statistical model: it measures the deviation between what one observes from data and what the forward model predicts (data fidelity term). One of the great advantages of working in a Banach space, is that the resulting methods are discretisation invariant and well-posed in the infinite-dimensional case. The minimisation problem (1.10) can be solved with standard optimisation techniques. Notice that from an algorithmic point of view, this is very similar to the computation of the MAP estimator as defined in (1.7). In fact, many of the Bayesian models that are frequently used in practice, can also be conceived from a variational perspective. In this light, the techniques that we establish in the thesis could be useful for variational models, particularly for problems where the variational formulation admits an interpretation from the lens of MAP estimation (e.g. denoising or deconvolution problems).

Finally, we want to discuss **statistical approaches**, which arise when at least one element in the inverse problem is modelled as stochastic. **Frequentist approaches** restrain the stochasticity to the elements that are perceived as intrinsically random, such as the noise, and then use traditional statistical tools to estimate the unknown parameters. **Bayesian methods** stem from a deeper level of abstraction: they conceive the inverse problem as a “statistical quest for information” [78], where all unknowns are treated as stochastic quantities, and then all available information is used to update the current state of knowledge about each quantity of interest. In this way, from a Bayesian point of view, the solution of an inverse problem is a posterior probability distribution, and not a single point estimate. If a point estimate is wanted, it can then be obtained from the posterior distribution by using decision theory [118]. In this context, the prior distributions used in Bayesian

models, play the same role as the frequentist penalty added in penalised maximum likelihood estimation (PMLE). Therefore, if we accept that the PMLE can be interpreted as a form of MAP estimation, then the proposed methodology can also be used for setting parameters in frequentist models.

## 1.7 Limitations of the existing approaches

Despite the breakthroughs in estimation accuracy and computing time, the existing methods for solving inverse problems are far from meeting the current needs and demands of the scientific community. While most modern mathematical imaging methods produce impressive point estimation results, they are generally unable to support the complex statistical analyses that are inherent to modern scientific reasoning.

First of all, most optimisation-only techniques struggle to deal with models when some of its parameters have unknown values. In particular, setting the regularisation parameters is notoriously difficult and this is the main area of contribution of this work. Although there are several general methods for setting regularisation parameters (see Chapter 2) most of them are limited to scalar regularisation parameters [3, 69, 101] or can only be applied to moderately ill-posed problems [93]. There are also many application-specific methods, which work well for particular problems but are hard to generalise [102, 129, 130]. Hence, the development of more general and robust tools for setting regularisation parameters remains an active research area [3, 49, 64, 105, 130], and is of great interest to the scientific imaging community [21, 32, 37, 38, 109].

Moreover, in applications related to quantitative imaging, where it is necessary to analyse images as high-dimensional physical measurements and not as pictures, obtaining point estimates alone is not enough. In these applications, it is essential to have an estimate of the uncertainty in the magnitudes of interest. Although most state-of-the-art methods do not quantify the uncertainty in the solutions they deliver, some recent works [20, 116] have adopted a Bayesian approach to address uncertainty quantification in imaging problems.

Modern scientific inquiry also requires advanced tools for selecting and comparing alternative mathematical models. Although there are some available methods for doing this when ground truth is available, intrinsic comparisons under no ground truth still remains a challenge. The heuristic for intrinsic model selection that we propose in this work is a contribution to this open problem.

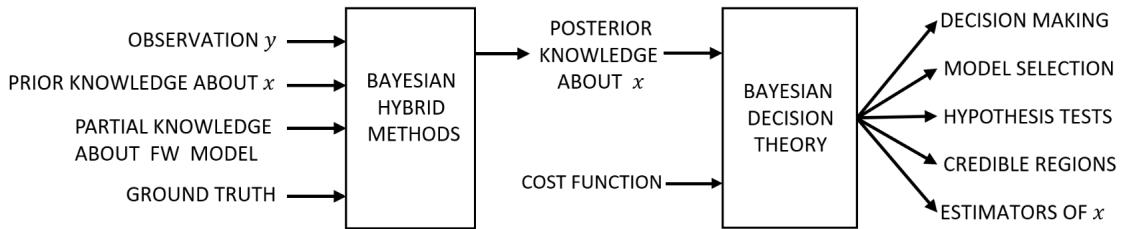
Furthermore, many modern disciplines use images as a mean for making decisions, which have associated costs. In this context, narrowing down the imaging

problems to the computation of a point estimate of the original image can sometimes leave out valuable information that would affect the final decision.

While none of the current approaches has yet managed to fully meet all these requirements, the Bayesian framework holds an enormous potential for supporting all these advanced statistical inquiries. When working under high-uncertainty scenarios this framework provides one of the most flexible and natural ways of integrating all sources of information and uncertainty under a single cohesive model.

Although pure Bayesian stochastic simulation for high-dimensional settings is sometimes too computationally expensive, recent works like [105], [56] and [109], show that there is a great potential for synergy between Bayesian and other non-statistical techniques.

Both neural networks and complex optimisation schemes can be plugged into a Bayesian backbone to create a new class of hybrid algorithms that are highly efficient, highly adaptive and still capable of performing sophisticated statistical inferences that are way beyond the scope of purely deterministic methods.



**Figure 1.5** – Bayesian hybrid methods for solving high-dimensional ill-posed should combine all available information (inputs) and output a summary of the current state of knowledge (the posterior distribution). This posterior information along with an optional cost function can be used to perform advanced statistical analysis underpinned by Bayesian Decision Theory.

Figure 1.5 shows a diagram illustrating the underlying structure of such Bayesian schemes. Here, for instance, data-driven techniques could be used to extract useful knowledge from the available ground truth. This knowledge could be combined with standard prior knowledge about  $x$  to construct a suitable regulariser. The observation  $y$  and all prior knowledge could be used to estimate the missing information about the forward model by, for example, using efficient optimisation schemes to compute point estimates of the unknown model parameters. All of this

combined, could then lead to some representation of the posterior knowledge about  $x$ , whether it is through an approximated surrogate model, or through simulated samples. Finally, this posterior knowledge about  $x$  could be used to carry out advanced statistical analysis.

On this account Bayesian techniques and, in particular, hybrid Bayesian methodology will probably be at the core of the next generation of computational imaging tools.

## 1.8 Publications

Different parts of this thesis have been accepted or are under review for publication in imaging and statistics journals:

- [136] A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part I: Methodology and Experiments*, to appear in SIAM Journal on Imaging Sciences (2020). Arxiv pre-print [137].
- [45] V. De Bortoli, D. Alain, M. Pereyra, and A. F. Vidal, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part II: Theoretical Analysis*, to appear in SIAM Journal on Imaging Sciences (2020). Arxiv pre-print [47].
- 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*, submitted to Statistics and Computing Springer Journal, currently under minor revision. Arxiv pre-print [46].
- A. F. Vidal, M. Pereyra, Giovannelli J.-F., *Fast model selection with empirical Bayesian priors*, in preparation.

Part of this work has also been presented at the 25th IEEE International Conference on Image Processing (ICIP) [138].

## 1.9 Other research activities

- Invited to give a talk *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach* at the Probability in the North-East (PiNE) Meeting, ICMS, Edinburgh, UK, Jan. 2020.
- Selected for oral presentation of *Maximum likelihood estimation of regularisation parameters: an empirical Bayesian approach* at the 2nd IMA Conference on Inverse Problems From Theory To Application, University College London, London, UK, Sep. 2019.
- Won best poster award presenting *Maximum likelihood estimation of regularisation parameters in imaging problems - an empirical Bayesian approach* at Annual PhD Poster Session, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK, Jun. 2019.
- Selected to go to Paris for three months to participate in the special semester *The Mathematics of Imaging* that took place at Institut Henri Poincaré, Paris, France, from Jan. 2019 to Apr. 2019. During these months I presented my work and collaborated with other researchers in France. In particular, I was invited to present my work at Neurospin and to collaborate with the Inria Parietal team to adapt my method to specific applications they work with.
- Attended The Mathematics of Imaging - Winter school and presented my work *Maximum likelihood estimation of regularisation parameters in imaging problems*, at Centre International de Rencontres Mathématiques, Marseille, France, Jan. 2019.
- Selected for oral presentation of *Maximum likelihood estimation of regularisation parameters* at the Statistical Signal Processing (SSP) Workshop 2018, STOR-i Centre for Doctoral Training, Lancaster University, Lancaster, UK, Apr. 2018.
- Gave a seminar *Maximum likelihood estimation of regularisation parameters in imaging inverse problems* as a part of the Actuarial Mathematics and Statistics Seminars, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK, Apr. 2018.

# Chapter 2

## Overview of existing methods for selecting regularisation parameters

In the following chapter we provide a brief overview of other existing methodologies for selecting regularisation parameters with a particular focus on strategies that are general enough to be applicable to a broad spectrum of problems. There are some specific techniques designed to work with particular models [9, 48], but we do not discuss them here.

### 2.1 Methods based on residual analysis

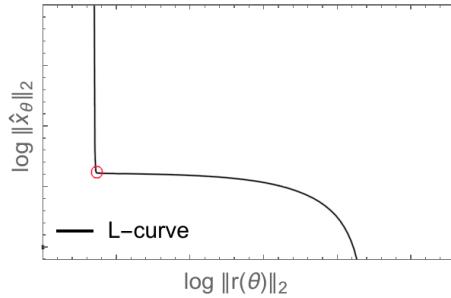
**Cross-validation** Cross-validation is a classic method for selecting regularisation parameters by analysing the residual  $r(\theta) = \|y - A\hat{x}_\theta\|$ , where  $\hat{x}_\theta$  is the solution to (1.7). This is a data-driven approach that proceeds as follows: i) split the pixels into  $n_g$  groups  $g_i$  (they may overlap) ii) for every pixel group  $g_i$  compute an estimate  $\hat{x}_\theta$  without using that group, iii) compute the pixel group residual  $r_i(\theta) = \|g_i - \hat{g}_i(\theta)\|$ , where  $\hat{g}_i(\theta)$  is given by extracting the same groups from  $A\hat{x}_\theta$  and iv) choose  $\theta_{CV}$  that minimises the cross-validatory function  $CV(\theta) = \sum_{\forall i} r_i(\theta)$  [128].

It is also possible to minimise a related function called the generalised cross-validatory function [65],  $GCV(\theta)$ , which does not differ much from the regular cross-validatory function, but has more desirable mathematical properties [69]: unlike  $\theta_{CV}$ , the parameter  $\theta_{GCV}$  that minimises  $GCV(\theta)$ , is invariant to orthogonal transformations and permutations of  $y$ . The GCV function can sometimes be quite

flat around the minimum resulting in numerical complications for finding its minimum [133].

Notice that both CV and GCV require solving the problem  $n_g$  times for every tested value of  $\theta$ , and this is too computationally expensive for most imaging problems.

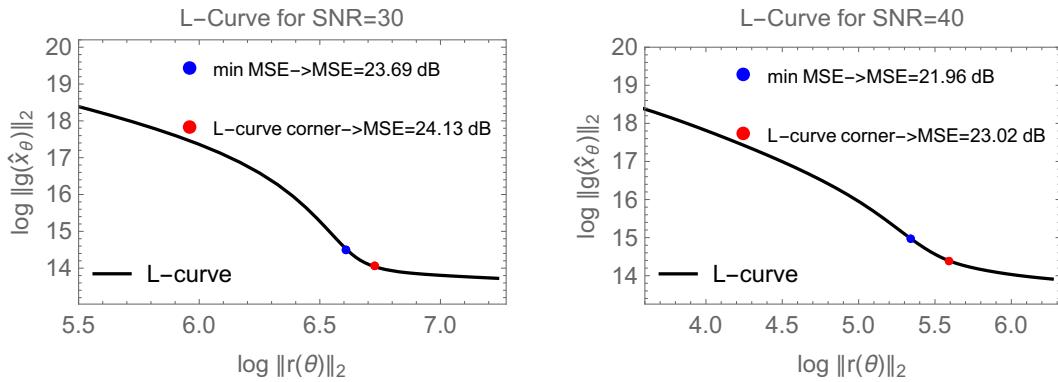
**The L-curve** The L-curve method is a popular alternative to CV and GCV that is more computationally efficient [69, 88]. The rationale behind this method is that when choosing the regularisation parameter, one should not only consider the norm of the residual, but also the norm of the regularisation term. To balance these two terms one can plot the L-curve, a parametric curve given by  $lc(\theta) = (\log \|r(\theta)\|_2, \log \|\hat{x}_\theta\|_2)$ , and then select the parameter  $\theta$  located at the corner of this curve (see Figure 2.1). This is supposed to be the point that better balances the different types of errors (coming from over-regularising or under-regularising).



**Figure 2.1** – Generic L-curve using Tikhonov regularisation. Example taken from [68].

As stated in [69], the performance of this method is close to the one of GCV, and it is not always a suitable option for imaging problems.

To illustrate some of the possible limitations, we show in Figure 2.2 an example of the L-curve for an image deblurring problem with a Total-Variation prior (as specified in Section 4.1.1). As it may be seen, the point that minimises the estimation mean squared error (MSE) does not necessarily fall in the corner of the L-curve. Moreover, the sharp corner seen on Figure 2.1 is not always present in many imaging problems. For higher SNR values the curve tends to flatten out making the estimation of the exact position of the corner more difficult and unstable. Finally, even if the optimal point is relatively close to the corner, we show that the



**Figure 2.2** – Deblurring of the `barbara` image with total-variation prior. L-curve for two different SNR values. Optimal point that minimises the reconstruction MSE in blue.

MSE is very sensitive to the exact position in the curve around the corner and using the L-curve criteria can lead to poor results.

**Discrepancy principle** The discrepancy principle is another popular approach that is often used in imaging problems [101]. Given a reliable estimate  $\varepsilon$  of the norm  $\|y - Ax\|$  for the true value of  $x$ , this method proposes to tune the regularisation parameter such that  $r(\theta) = \|y - A\hat{x}_\theta\|$  is close to  $\varepsilon$ . This is based on the intuitive notion that when the computed solution is sufficiently regularised, the residual should be primarily constituted of noise.

For models where  $y \in \mathbb{C}^{d_y}$  is given by  $Ax + w$  with  $w \sim \mathcal{N}(0, \sigma^2 I_{d_y})$ , the norm of the noise  $\|w\|_2^2$  follows a  $\sigma^2 \chi^2(d_y)$  distribution with mean  $\mu_{\chi^2} = \sigma^2 d_y$  and variance  $\sigma_{\chi^2}^2 = 2 d_y \sigma^4$ . A common choice for  $\varepsilon$  is therefore  $\varepsilon = \mu_{\chi^2}$  or  $\varepsilon = \mu_{\chi^2} + 2\sigma_{\chi^2}$  (when  $d_y$  is large the difference is not very significant). Given this bound,  $\theta$  can be obtained by either finding  $\theta_{DP}$  such that  $r(\theta_{DP}) = \varepsilon$ , or alternatively  $\theta_{DP} = \operatorname{argmax}_{\theta \in \Theta} r(\theta)$  s.t.  $r(\theta) \leq \varepsilon$ .

A main limitation of this approach is that it can only be applied when  $\theta$  is scalar, and that it requires having a good estimate of the noise level (which is not always available). Furthermore, it tends to overestimate the regularisation parameter leading to solutions that are over-smoothed, especially in cases of low signal to noise ratio (SNR) [69]. This phenomenon can be observed in some of the numerical experiments presented in section 4.1. This approach is also sensitive to model misspecification, as it heavily relies on the likelihood to set  $\theta$ .

**Residual whiteness measures** A more recent method within this group is the one proposed in [3], where they set the regularisation parameter in a blind deconvolution problem by minimising the autocorrelation of the residual. Although this is an interesting alternative for this kind of problem, it is not optimal for non-blind problems where information about the operator  $A$  is available.

Residual based methods have been mostly superseded by newer approaches such as the ones we discuss in the following two subsections [49, 105]. Since the discrepancy principle is still frequently used in imaging inverse problems [1, 31], we will compare it with the proposed method in the cases where it is possible (when  $\theta$  is scalar).

## 2.2 Methods based on surrogates of the MSE

If we could compute the estimation error (MSE) for each value of  $\theta$ , we could choose the value of  $\theta$  that minimises this error. This is of course not possible because we do not have access to the true image. However, it is sometimes possible to construct an estimator of this estimation error that does not depend on the unknown underlying image, and then minimise this estimator instead.

This is the approach adopted in the Stein’s Unbiased Risk Estimator (SURE) [127] methods, which have recently received a lot of attention [49, 57, 64, 110, 114]. Although SURE was originally conceived for denoising problems with white additive Gaussian noise, works like [57] have extended the results for more general inverse problems (not only denoising) and for a wider class of noise distributions within the exponential family. This generalised SURE (GSURE) however, is not enough to tackle the cases where the operator  $A$  has a non-trivial null space. GSURE uses an estimator of  $\|x - \hat{x}_\theta\|_2^2$  which can be computed provided  $A^\top A$  is invertible. When  $A^\top A$  is not invertible, there is a significant difference between minimizing the *estimation* risk  $E\{\|x - \hat{x}_\theta\|_2^2\}$  and the *prediction* risk  $E\{\|Ax - A\hat{x}_\theta\|_2^2\}$ . In [64] they propose the alternative Projected-GSURE, which uses an estimate of the *projection* risk which considers the error computed only on the projection of  $x - \hat{x}_\theta$  onto the range of  $A^\top$ . In this way only the components of  $x$  that can be “observed” through  $A$  are taken into account, making the projection risk a better approximation of the

estimation risk for general ill-posed problems.

Traditionally, SURE-based methods were carried out by exhaustive search, trying out different values for  $\theta$ . More advanced methods such as SUGAR [49] find an asymptotically unbiased estimate of the gradient of SURE, and then use this to find  $\theta$  with an optimisation scheme.

Despite of the attempts to extend the SURE-based methodology for more ill-posed inverse problems, it still faces some major limitations when it comes to severely ill-posed problems [93]. Most of the published works [28, 49, 64, 142] that promote the use of this technique for parameter selection actually considered only very “mildly ill-posed” problems. A very recent work [93] has studied the limitations of SURE-methods in depth and argues that these techniques do not constitute a reliable approach for general ill-posed problems. In this Section 4.1.1 we compare our method to the one proposed in [49] and we observe the same behaviour reported in [93], namely that the regularisation parameter is underestimated.

## 2.3 Bayesian methods

As we mentioned earlier, the Bayesian framework provides two main paradigms to select  $\theta$  automatically: the empirical (discussed in section 2.3.2) and the hierarchical [96] (discussed in section 2.3.1).

### 2.3.1 Hierarchical Bayesian estimation

In the hierarchical paradigm,  $\theta$  is modelled as an additional unknown quantity and it is assigned a prior distribution  $p(\theta)$ . This leads to an augmented posterior given by

$$p(x, \theta | y) = \frac{p(y|x, \theta)p(x|\theta)p(\theta)}{p(y)}. \quad (2.1)$$

There are two main ways of employing this augmented posterior. One possibility is to estimate  $x$  and  $\theta$  jointly from  $y$  [97]. For example, one can perform maximum-a-posteriori estimation jointly on  $x$  and  $\theta$ , i.e.

$$(\hat{x}_*, \hat{\theta}_*) = \underset{x \in \mathbb{R}^d, \theta \in \Theta}{\operatorname{argmax}} p(x, \theta | y). \quad (2.2)$$

Although this approach has been successfully applied in specific problems such as [146], obtaining a general implementation is difficult because maximising  $p(x, \theta|y)$  requires knowing  $Z(\theta)$ , the normalising constant of  $p(x|\theta)$  defined in (1.3). As mentioned earlier, in most models of interest (1.3) is intractable, so developing a general algorithm to compute (2.2) can be very challenging.

Alternatively, given that in imaging problems the main goal is to recover  $x$  and the actual value of  $\theta$  is not relevant, the most popular approach is to remove  $\theta$  from the model by marginalisation followed by inference on  $x|y$  with the marginal posterior given by

$$p(x|y) = \int_{\Theta} p(x, \theta|y) d\theta. \quad (2.3)$$

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

$$\hat{x}_{\text{MMSE}} = \int_{\mathbb{R}^d} \tilde{x} p(\tilde{x}|y) d\tilde{x}. \quad (2.4)$$

This can be achieved with a standard MCMC algorithm when  $Z(\theta)$  is tractable, e.g. Gibbs sampling, or with specialised algorithms that allow to circumvent the evaluation of  $Z(\theta)$  at the expense of significant additional computational cost (see [106] 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 [9, 94]), but such algorithms have not yet been widely adopted because their implementation and performance remains very problem-specific. Alternatively, one can also compute the marginal MAP estimator

$$\hat{x}_{\text{MAP}} \in \operatorname{argmin}_{\tilde{x} \in \mathbb{R}^d} p(\tilde{x}|y), \quad (2.5)$$

which, unlike the MMSE estimator and other Bayesian estimators, can be computed with optimisation algorithms.

In particular, for log-concave posteriors, both (2.5) and (2.2) can be obtained using the efficient majorisation-minimisation algorithms proposed in [105]. This recent work introduces an ingenious way of using the exact normalising constant  $Z(\theta)$ , even when it is not available in an explicit form: the authors show that for a

$\alpha$ -homogeneous regulariser<sup>1</sup>  $g(x)$  and a scalar  $\theta$ , the normalising constant is always of the form

$$Z(\theta) = c \theta^{-n/\alpha} \quad (2.6)$$

where  $c$  is a constant that does not depend on  $\theta$ . Based on this insight, [105] proposes specific algorithms to compute  $(\hat{x}_*, \hat{\theta}_*)$  and  $\hat{x}_{MAP}$  for models with homogeneous regularisers and using the prior distribution

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbb{1}_{\mathbb{R}^+}(\theta), \quad (2.7)$$

with  $a \in \mathbb{R}^+$  and  $b \in \mathbb{R}^+$ .

Since the algorithms proposed in [105] are suitable for some of the high dimensional imaging problems we consider in this thesis, we use it as benchmark in some experiments. Although the hierarchical paradigm is theoretically compatible with multivariate regularisation parameters, the implementation offered in [105] is mainly useful for scalar parameters and homogeneous regularisers, and can only be applied to a multivariate  $\theta$  for very specific cases where either the prior distribution is multiplicatively separable (see remark 3.1.2) or the normalising constant of the prior,  $Z(\theta)$ , is known.

Lastly, it is worth mentioning that the hierarchical strategy has been studied in detail in [24, 25] in the context of hierarchical Bayesian sparse regularisation models. More precisely, these works cleverly exploit the conditional structure of certain hierarchical Gaussian models with random prior covariance matrices to propose a simple iterative alternating scheme to compute the joint MAP estimator of  $x$  and the prior covariance. This kind of scheme yields good results for the class of imaging models in those works, both in terms of accuracy and computational complexity. The generalisation of the ideas of [24, 25] to other imaging models, particularly the class of models considered in this thesis, is very interesting but highly non-trivial as the approach they use to analytically decompose the marginal likelihood relies heavily on the Gaussian nature of the model and cannot be generalised in a straightforward way.

---

<sup>1</sup> $g$  is a  $\alpha$ -homogeneous function if there exists  $\alpha \in \mathbb{R}^+$  such that  $g(\lambda x) = \lambda^\alpha g(x) \forall x \in \mathbb{R}^d, \forall \lambda > 0$

### 2.3.2 Empirical Bayes estimation

Under an empirical Bayesian paradigm, the regularisation parameter  $\theta \in \Theta$  is estimated directly from the observed data  $y$ . The term *empirical* refers to the fact that part of the prior distribution (in this case the hyperparameter  $\theta$ ) is estimated from the observation instead of being fully specified a priori. In this thesis we adopt this approach and, in particular, we compute  $\theta_*$  by maximum marginal likelihood estimation, that is,

$$\theta_* \in \operatorname{argmax}_{\theta \in \Theta} p(y|\theta), \quad (2.8)$$

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

$$p(y|\theta) = \int_{\mathbb{R}^d} p(y|\tilde{x})p(\tilde{x}|\theta)d\tilde{x}. \quad (2.9)$$

Given  $\theta_*$ , empirical Bayesian approaches base inferences on the pseudo-posterior  $x \mapsto p(x|y, \theta_*)$  [30] given by

$$p(x|y, \theta_*) = \exp[-f_y(x) - \theta_* g(x)] \Big/ \int_{\mathbb{R}^d} \exp[-f_y(\tilde{x}) - \theta_* g(\tilde{x})]d\tilde{x}. \quad (2.10)$$

Observe that this strategy is equivalent to Bayesian model selection on a continuous class of models parametrised by  $\theta$ , where  $\theta_*$  produces the model with the best fit-to-data (under some additional assumptions,  $p(y|\theta_*)$  provides the best approximation of the true distribution of  $y$  in a Kullback–Leibler divergence sense [143]).

Empirical Bayesian approaches were first considered in the statistical methodology community (see e.g. [30, 117]), which stimulated developments in computational statistics [6, 7, 119] 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. [82, 111, 123]).

Unfortunately, this powerful inference strategy is difficult to apply in imaging problems [124] because the marginal likelihood  $\theta \mapsto p(y|\theta)$  is computationally intractable as it involves two  $d$ -dimensional integrals, namely (1.3) and (2.9), thus making the optimisation problem (2.8) very challenging. As mentioned previously, the aim of this thesis 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 [6, 7], we propose a stochastic gradient MCMC algorithm to efficiently solve (2.8) for imaging models of the general form (1.6), where two main novelties are that we use state-of-the-art proximal MCMC methods [56] to construct a stochastic optimisation scheme that scales efficiently to high dimensions, and that we provide easily verifiable theoretical conditions ensuring convergence.

The maximum likelihood estimation problem (2.8) raises natural questions about the uniqueness of  $\theta_*$ , 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  $\theta_*$ . 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  $\theta_*$ , and is strongly log-concave w.r.t.  $\theta$  in the neighbourhood of  $\theta_*$ . These favourable properties can be formally derived under simplifying assumptions (e.g. that  $p(y|\theta)$  is fully separable on  $y$  [132]). Extending conditions for uniqueness of (2.8) to more general imaging problems is an important perspective for future work.

Lastly, we note that empirical Bayesian methods have found many applications in machine learning, for example in the context of feature selection (see, e.g., [102, 129, 130]). In this field, the challenges related to high-dimensionality have been mainly addressed by using conditional Gaussian models for which the high-dimensional integrals (1.3) and (2.9) become tractable, thus enabling the use of specialised strategies to solve the optimisation problem (2.8).

### 2.3.3 Connections between both Bayesian approaches

As we mentioned earlier, the Bayesian framework provides two main paradigms to select  $\theta$  automatically: the empirical and the hierarchical, which is currently the predominant Bayesian approach in data science (see [105, 106] for examples in imaging sciences). We now discuss connections between the two paradigms and stress advantages and disadvantages.

In order to understand the connection between this hierarchical Bayesian ap-

proach and the empirical Bayesian strategy used in this thesis it is useful to express  $p(x|y)$  as follows

$$p(x|y) = \int_{\Theta} p(x|y, \tilde{\theta}) p(\tilde{\theta}|y) d\tilde{\theta}, \quad (2.11)$$

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  $\theta$  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  $\delta_{\theta_*}$ , we obtain the empirical Bayesian pseudo-posterior  $x \mapsto p(x|y, \theta_*)$  considered in this manuscript.

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  $\theta_*$ . 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 [118].

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, what we observe in practice is that the marginal posterior  $p(\theta|y)$  typically has its maximum at a good value for  $\theta$ , but struggles to concentrate and spreads its mass across a much wider range of values of  $\theta$ . 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  $x \mapsto p(x|y, \theta_*)$  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 6.4, and is particularly clear in the hyperspectral unmixing problem.

# Chapter 3

## Proposed methodology

In this chapter, we present our new empirical Bayesian method for setting regularisation parameters in problems where the posterior distribution is log-concave, i.e. where maximum-a-posteriori estimation is a convex optimisation problem. We estimate the regularisation parameters directly from the observed data by maximum marginal likelihood estimation. To maximise the intractable marginal likelihood, we use a stochastic proximal gradient algorithm that is driven by non-homogeneous Markov chain samplers.

We start the chapter by introducing three alternative algorithms: the first two use a single Markov kernel and can be used in problems with homogeneous regularisers, or where the log-prior has tractable derivatives, as this allows evaluating the required prior expectations explicitly. The third algorithm is more general, as it does not require the explicit evaluation of derivatives of the log-prior but rather approximates the required prior expectations by using an additional Markov kernel targeting the prior distribution of  $x$ . In Section 3.1.4, we provide details about the Markov kernels used to drive the stochastic gradient algorithm and in Section 3.1.5 we discuss connections between the proposed methodology and the expectation-maximisation algorithm.

In Section 3.2, we illustrate the method by considering a simple image denoising problem, where we work with synthetic test images for which the exact generative statistical model is known. This allows assessing the performance of the method in a case where the regularisation parameter has a true value, and where there is no model misspecification. We use this example to i) show the role of different model

parameters, ii) study the statistical behaviour of the method under both Gaussian and Laplace noise, iii) test the performance under extreme noise conditions and iv) explore the robustness of the method towards mild likelihood misspecification (e.g., when there is a mismatch in the statistical properties of the noise).

In Section 3.3 we provide very comprehensive implementation guidelines, where we explain how to set all the algorithm parameters and include other implementation and troubleshooting recommendations.

### 3.1 Proposed algorithm

We now present the proposed empirical Bayesian method to solve the marginal maximum likelihood estimation problem (2.8) and set regularisation parameters. As mentioned previously, the main difficulty in solving (2.8) 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  $\Theta$  is a convex compact set, we could seek to iteratively solve (2.8) by using the projected gradient algorithm [42] which is given by  $(\theta_n)_{n \in \mathbb{N}}$  with  $\theta_0 \in \Theta$  and associated with the following recursion for any  $n \in \mathbb{N}$

$$\theta_{n+1} = \Pi_\Theta [\theta_n + \delta_{n+1} \nabla_\theta \log p(y|\theta_n)] , \quad (3.1)$$

where  $\Pi_\Theta$  is the projection onto  $\Theta$  and  $(\delta_n)_{n \in \mathbb{N}}$  is a sequence of non-increasing step-sizes. 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  $\theta_*$  and is strongly log-concave w.r.t.  $\theta$  in the neighbourhood of  $\theta_*$ . Therefore we expect that (3.1) would quickly converge.

Since  $\theta \mapsto \nabla_\theta \log p(y|\theta)$  is not tractable, we cannot directly use (3.1) to compute  $\theta_*$ . However, we can replace  $\theta \mapsto \nabla_\theta \log p(y|\theta)$  with a noisy estimate and consider a stochastic variant of the projected gradient algorithm. In particular, under mild assumptions using Fisher's identity (see Proposition 1 in Appendix A) 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

$$\theta \in \Theta$$

$$\nabla_\theta \log p(y|\theta) = \int_{\mathbb{R}^d} p(\tilde{x}|y, \theta) \nabla_\theta \log p(\tilde{x}, y|\theta) d\tilde{x} = - \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|y, \theta) d\tilde{x} - \nabla_\theta \log Z(\theta) . \quad (3.2)$$

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 [60], where the expectation  $\int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|y, \theta) d\tilde{x}$  is replaced by a Monte Carlo estimator leading to the following gradient estimate for any  $\theta \in \Theta$

$$\Delta_{m,\theta} = \frac{1}{m} \sum_{k=1}^m \nabla_\theta \log p(X_k, y|\theta) = -\nabla_\theta \log Z(\theta) - \frac{1}{m} \sum_{k=1}^m g(X_k) , \quad (3.3)$$

where  $(X_k)_{k \in \{0, \dots, m\}}$  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, to compute  $\theta_*$ , we can build a new sequence  $(\theta_n)_{n \in \mathbb{N}}$  associated with the following recursion for any  $n \in \mathbb{N}$

$$\theta_{n+1} = \Pi_\Theta[\theta_n + \delta_{n+1} \Delta_{m_n, \theta_n}] , \quad \Delta_{m_n, \theta_n} = -\nabla_\theta \log Z(\theta_n) - \frac{1}{m_n} \sum_{k=1}^{m_n} g(X_k^n) , \quad (3.4)$$

starting from some  $\theta_0 \in \Theta$ , and where  $(m_n)_{n \in \mathbb{N}}$  is a sequence of non-decreasing sample sizes. Under some assumptions on  $(m_n)_{n \in \mathbb{N}}$ ,  $(\delta_n)_{n \in \mathbb{N}}$  and on the Markov kernels (see Theorem 6 in Appendix C), 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  $N \in \mathbb{N}$ , a sequence of non-increasing weights  $(\omega_n)_{n \in \mathbb{N}}$ , and a sequence  $(\theta_n)_{n=0}^{N-1}$  generated using (3.4), an approximate solution of (2.8) can be obtained by calculating, for example, the weighted average<sup>1</sup>

$$\bar{\theta}_N = \sum_{n=0}^{N-1} \omega_n \theta_n \left/ \sum_{n=0}^{N-1} \omega_n \right. , \quad (3.5)$$

which converges asymptotically to a solution of (2.8) as  $N \rightarrow \infty$  (see [7] for details).

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

---

<sup>1</sup>Averaging iterates is standard in stochastic approximation algorithms. Most known convergence results concern the almost sure convergence of  $(p(y|\bar{\theta}_N))_{N \in \mathbb{N}}$  towards  $\min_{\theta \in \Theta} p(y|\theta)$ , or alternatively a weaker convergence in expectation (see, e.g., [7, 11, 113]).

a way that is computationally efficient and that satisfies a number of complex technical conditions on the associated Markov kernels (see Theorem 6 in Appendix C). In this work, we address this major difficulty by constructing an SAPG scheme with state-of-the-art unadjusted proximal Markov kernels  $\{R_{\gamma,\lambda,\theta} : \gamma \in (0, \bar{\gamma}], \lambda \in \mathbb{R}^+, \theta \in \Theta\}$  that automatically satisfy the required theoretical conditions. Here  $\gamma$  and  $\lambda$  are kernel parameters that control a trade-off between accuracy and computational efficiency (we provide more details about the kernels in Section 3.1.4). More importantly, we show both theoretically and empirically that a single sample ( $m_n = 1$ ) per iteration is enough to guarantee the convergence of the proposed SAPG scheme. This allows delivering accurate estimates of regularisation parameters in a computationally scalable way and with theoretical guarantees.

Lastly, observe that in order to use (3.4) it is necessary to evaluate  $\theta \mapsto \nabla_\theta \log Z(\theta)$ . For most models of interest,  $\theta \mapsto \nabla_\theta \log Z(\theta)$  cannot be computed exactly and needs to be approximated. Hence, we propose three different strategies to address this calculation depending on whether  $g$  is a homogeneous function or not.

### 3.1.1 Scalar-valued $\theta$ with homogeneous regulariser

Let  $g$  be a homogeneous regulariser of degree  $\alpha \in \mathbb{R} \setminus \{0\}$ , i.e. for any  $x \in \mathbb{R}^d$  and  $t > 0$ ,  $g(tx) = t^\alpha g(x)$ , then for scalar-valued  $\theta$ , i.e.  $d_\Theta = 1$ , (3.2) is given by

$$\frac{d}{d\theta} \log p(y|\theta) = - \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|y, \theta) d\tilde{x} - \frac{d}{d\theta} \log Z(\theta). \quad (3.6)$$

Recalling that  $\Theta \subset (0, +\infty)$  we have for any  $\theta \in \Theta$

$$Z(\theta) = \int_{\mathbb{R}^d} e^{-\theta g(\tilde{x})} d\tilde{x} = \int_{\mathbb{R}^d} e^{-g(\theta^{1/\alpha} \tilde{x})} d\tilde{x} = \theta^{-d/\alpha} \int_{\mathbb{R}^d} e^{-g(\tilde{x})} d\tilde{x}, \quad (3.7)$$

and therefore

$$\frac{d}{d\theta} \log Z(\theta) = -d/(\alpha\theta). \quad (3.8)$$

Hence, (3.6) becomes for any  $\theta \in \Theta$

$$\frac{d}{d\theta} \log p(y|\theta) = d/(\alpha\theta) - \int_{\mathbb{R}^d} g(\tilde{x}) p(\tilde{x}|y, \theta) d\tilde{x}, \quad (3.9)$$

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  $\ell_1$ ,  $\ell_2$ , 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  $\alpha$ , are also  $\alpha$  positively homogeneous. Notice that Algorithm **1** does not require sampling from the prior distribution, so it admits the use of improper priors [118, Section 1.5] such as total-variation.

---

**Algorithm 1** SAPG algorithm - Scalar  $\theta$  and  $\alpha$  positively homogeneous regulariser  $g$

---

- 1: **Input:** initial  $\{\theta_0, X_0^0\}$ ,  $(\delta_n, \omega_n, m_n)_{n \in \mathbb{N}}$ ,  $\Theta$ , kernel parameters  $\gamma, \lambda$ , iterations  $N$ .
- 2: **for**  $n = 0$  to  $N - 1$  **do**
- 3:     **if**  $n > 0$  **then**
- 4:         Set  $X_0^n = X_{m_{n-1}}^{n-1}$ ,
- 5:     **end if**
- 6:     **for**  $k = 0$  to  $m_n - 1$  **do**
- 7:         Sample  $X_{k+1}^n \sim R_{\gamma, \lambda, \theta_n}(X_k^n, \cdot)$ ,
- 8:     **end for**
- 9:     Set  $\theta_{n+1} = \Pi_\Theta \left[ \theta_n + \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \left\{ \frac{d}{\alpha \theta_n} - g(X_k^n) \right\} \right]$ .
- 10: **end for**
- 11: **Output:**  $\bar{\theta}_N$  computed with (3.5).

---

In Section 3.1.5, we draw connections between the proposed method and the expectation-maximisation algorithm, and then use those connections to propose a variant of Algorithm **1** and Algorithm **2** where the gradient step ( Line 9 in Algorithm **1** and Line 5 in Algorithm **2**) is replaced by a full maximisation in closed form.

### 3.1.2 Separably homogeneous regulariser

For the special case of separably homogeneous regularisers, Algorithm **1** can be adapted for multivariate  $\theta$ . This is because in this class of regulariser, each component of  $\theta$  affects independent subsets of the components of  $x$ . More precisely, assume that  $g$  is separably homogeneous in the following sense: there exist  $(\tilde{g}_i)_{i \in \{1, \dots, d_\Theta\}}$ ,  $(A_i)_{i \in \{1, \dots, d_\Theta\}}$  pairwise disjoint subsets of  $\{1, \dots, d\}$  and  $(\alpha_i)_{i \in \{1, \dots, d_\Theta\}}$  such that for

any  $i \in \{1, \dots, d_\Theta\}$ ,  $\tilde{g}_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}$  is  $\alpha_i$ -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  $\theta \in \Theta$

$$Z(\theta) = \int_{\mathbb{R}^d} \exp[-\theta^\top g(\tilde{x})] d\tilde{x} = \int_{\mathbb{R}^d} \exp \left[ -\sum_{i=1}^{d_\Theta} \theta^i \tilde{g}_i(\tilde{x}_{[\mathsf{A}_i]}) \right] d\tilde{x} \quad (3.10)$$

$$= \prod_{i=1}^{d_\Theta} \int_{\mathbb{R}^{|\mathsf{A}_i|}} \exp[-\theta^i \tilde{g}_i(\tilde{x}_{[\mathsf{A}_i]})] d\tilde{x}. \quad (3.11)$$

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

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

Using this property we obtain Algorithm 2, where for any  $i \in \{1, \dots, d_\Theta\}$ ,  $\theta^i \in \Theta^i \subset (0, +\infty)$  and  $\Pi_{\Theta^i}$  is the projection onto  $\Theta^i$ .

---

**Algorithm 2** SAPG algorithm - Multivariate  $\theta$  and separably homogeneous regulariser

---

- 1: **Input:** initial  $\{\theta_0, X_0^0\}$ ,  $(\delta_n, \omega_n, m_n)_{n \in \mathbb{N}}$ ,  $\Theta$ , kernel parameters  $\gamma, \lambda$ , iterations  $N$ .
  - 2: **for**  $n = 0$  to  $N - 1$  **do**
  - 3:     **if**  $n > 0$  **then**
  - 4:         Set  $X_0^n = X_{m_{n-1}}^{n-1}$ ,
  - 5:     **end if**
  - 6:     **for**  $k = 0$  to  $m_n - 1$  **do**
  - 7:         Sample  $X_{k+1}^n \sim R_{\gamma, \lambda, \theta_n}(X_k^n, \cdot)$ ,
  - 8:     **end for**
  - 9:     **for**  $i = 1$  to  $d_\Theta$  **do**
  - 10:         Set  $\theta_{n+1}^i = \Pi_{\Theta^i} \left[ \theta_n^i + \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \left\{ \frac{|\mathsf{A}_i|}{\alpha_i \theta_n^i} - \tilde{g}_i(X_{k[\mathsf{A}_i]}^n) \right\} \right]$ .
  - 11:     **end for**
  - 12: **end for**
  - 13: **Output:**  $\bar{\theta}_N$  computed with (3.5).
- 

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<sup>2</sup>, and consider level-adapted  $\ell_1$  regularisations of

---

<sup>2</sup>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  $\Psi^\top x$ .

the form

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

where  $x_{[\mathbf{A}_j]}$  are the elements of  $x$  associated with the  $J$ th 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  $d_j = |\mathbf{A}_j|$ . 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  $Z(\theta)$  holds.

### 3.1.3 General case: inhomogeneous regulariser

When  $g$  is neither homogeneous nor separably homogeneous, we address the evaluation of  $\theta \mapsto \nabla_\theta \log Z(\theta)$  numerically by stochastic simulation. More precisely, using identity (1.4) we can express the intractable term  $-\nabla_\theta \log Z(\theta)$  as an expectation  $\int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta) d\tilde{x}$ , which we can also replace with a Monte Carlo estimate by using an additional Markov kernel that samples from the prior  $p(x|\theta)$ . In this way, using that  $y$  is conditionally independent of  $\theta$  given  $x$ , we can rewrite  $\theta \mapsto \nabla_\theta \log p(y|\theta)$  as the difference between two expectations, *i.e.* for any  $\theta \in \Theta$

$$\nabla_\theta \log p(y|\theta) = \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta) d\tilde{x} - \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|y, \theta) d\tilde{x}, \quad (3.12)$$

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

### 3.1.4 MCMC Kernels

Given the high dimensionality involved, it is fundamental to carefully choose the families of Markov kernels  $\{\mathbf{R}_{\gamma, \lambda, \theta}, \bar{\mathbf{R}}_{\gamma', \lambda', \theta} : \gamma, \gamma' \in (0, \bar{\gamma}], \lambda, \lambda' \in \mathbb{R}^+, \theta \in \Theta\}$  driving the SAPG. Here we use the Moreau-Yosida Unadjusted Langevin Algorithm (MYULA) Markov kernel recently proposed in [56], 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 particu-

---

**Algorithm 3** SAPG algorithm - General form

---

```

1: Input: initial  $\{\theta_0, X_0^0, \bar{X}_0^0\}$ ,  $(\delta_n, \omega_n, m_n)_{n \in \mathbb{N}}$ ,  $\Theta, \gamma, \gamma', \lambda, \lambda'$ , iterations  $N$ .
2: for  $n = 0$  to  $N - 1$  do
3:   if  $n > 0$  then
4:     Set  $X_0^n = X_{m_{n-1}}^{n-1}$ ,
5:     Set  $\bar{X}_0^n = \bar{X}_{m_{n-1}}^{n-1}$ ,
6:   end if
7:   for  $k = 0$  to  $m_n - 1$  do
8:     Sample  $X_{k+1}^n \sim R_{\gamma, \lambda, \theta_n}(X_k^n, \cdot)$ ,
9:     Sample  $\bar{X}_{k+1}^n \sim \bar{R}_{\gamma', \lambda', \theta_n}(\bar{X}_k^n, \cdot)$ ,
10:   end for
11:   Set  $\theta_{n+1} = \Pi_\Theta \left[ \theta_n + \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \{g(\bar{X}_k^n) - g(X_k^n)\} \right]$ .
12: end for
13: Output:  $\bar{\theta}_N$  computed with (3.5).

```

---

lar MCMC method is derived from the discretisation of an over-damped Langevin diffusion,  $(\bar{X}_t)_{t \geq 0}$ , satisfying the following stochastic differential equation

$$d\mathbf{X}_t = -\nabla_x F(\mathbf{X}_t)dt + \sqrt{2}dB_t, \quad (3.13)$$

where  $F : \mathbb{R}^d \mapsto \mathbb{R}$  is a continuously differentiable potential and  $(B_t)_{t \geq 0}$  is a standard  $d$ -dimensional Brownian motion. Under mild assumptions, this equation has a unique strong solution [74, Chapter 4, Theorem 2.3]. Accordingly, the law of  $(X_t)_{t \geq 0}$  converges as  $t \rightarrow \infty$  to the diffusion's unique invariant distribution, with probability density given by  $\pi(x) \propto e^{-F(x)}$  for all  $x \in \mathbb{R}^d$  [121, Theorem 2.2]. Hence, to use (3.13) as a Monte Carlo method to sample from the posterior  $p(x|y, \theta)$ , we set  $F(x) = -\log p(x|y, \theta)$  and thus specify the desired target density. Similarly, to sample from the prior we set  $F(x) = -\log p(x|\theta)$ .

However, sampling directly from (3.13) is usually not computationally feasible. Instead, we usually resort to a discrete-time Euler-Maruyama approximation<sup>3</sup> [55] of (3.13) 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

$$X_{k+1} = X_k - \gamma \nabla_x F(X_k) + \sqrt{2\gamma} Z_{k+1}, \quad (3.14)$$

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 ma-

---

<sup>3</sup>The Euler-Maruyama approximation is a generalisation of the forward Euler method for ordinary differential equations to stochastic differential equations.

trix. This Markov chain is commonly known as the Unadjusted Langevin Algorithm (ULA) [121]. Under some additional assumptions on  $F$ , namely Lipschitz continuity of  $\nabla_x F$ , the ULA chain inherits the convergence properties of (3.13) and converges to a stationary distribution that is close to the target  $\pi$ , with  $\gamma$  controlling a trade-off between accuracy and convergence speed [56].

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  $V(x)$ , see [13, Definition 12.20], defined for any  $x \in \mathbb{R}^d$  and  $\lambda > 0$  by

$$V^\lambda(x) = \min_{\tilde{x} \in \mathbb{R}^d} \{V(\tilde{x}) + (1/2\lambda) \|x - \tilde{x}\|_2^2\} , \quad (3.15)$$

where one can make  $V^\lambda(x)$  arbitrarily close to  $V(x)$  by reducing the smoothing parameter  $\lambda$  (see [56] for details). For any  $\lambda > 0$ , the Moreau-Yosida envelope  $V^\lambda$  is continuously differentiable with gradient given for any  $x \in \mathbb{R}^d$  by

$$\nabla V^\lambda(x) = (x - \text{prox}_V^\lambda(x))/\lambda , \quad (3.16)$$

(see, e.g., [13, 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  $k \in \mathbb{N}$

$$\text{MYULA} : X_{k+1} = X_k - \gamma \nabla_x U(X_k) - \gamma \nabla_x V^\lambda(X_k) + \sqrt{2\gamma} Z_{k+1} . \quad (3.17)$$

Returning to the problem of interest, if we define the splitting such that  $U = f_y$  and

$V = \theta^\top g$ , we can define the MYULA families of Markov kernels  $\{R_{\gamma,\lambda,\theta}, \bar{R}_{\gamma',\lambda',\theta} : \gamma, \gamma' \in (0, \bar{\gamma}], \lambda, \lambda' \in \mathbb{R}^+, \theta \in \Theta\}$  that we use in Algorithm 1, Algorithm 2 and Algorithm 3. For any  $\theta \in \Theta$ ,  $\gamma > 0$  and  $\lambda > 0$ , the kernel  $R_{\gamma,\lambda,\theta}$  associated with  $(X_k)_{k \in \mathbb{N}}$  starting from  $X_0 \in \mathbb{R}^d$ , is given by the following recursion for any  $k \in \mathbb{N}$

$$X_{k+1} = X_k - \gamma \nabla_x f_y(X_k) - \gamma \{X_k - \text{prox}_{\theta^\top g}^\lambda(X_k)\} / \lambda + \sqrt{2\gamma} Z_{k+1}. \quad (3.18)$$

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

$$\bar{X}_{k+1} = \bar{X}_k - \gamma' \{\bar{X}_k - \text{prox}_{\theta^\top g}^{\lambda'}(\bar{X}_k)\} / \lambda' + \sqrt{2\gamma'} Z_{k+1}, \quad (3.19)$$

where we recall that  $\lambda$  and  $\lambda'$  are the smoothing parameters associated with  $\theta^\top g^\lambda$ ;  $\gamma$  and  $\gamma'$  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.

**Alternative implementations of MYULA** 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 (3.18) 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  $g_i^\lambda$  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 (1.7) also use the operators  $\nabla f_y$  and either  $\text{prox}_{\theta^\top g}^\lambda$  or  $\text{prox}_{\theta^\top g_i}^\lambda$  [42, 66], making the implementation of the proposed methodology straightforward for problems that are currently solved with such tools.

**Estimation bias** We note at this point that the MYULA kernels (3.18) and (3.19), 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 (3.13), 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  $\lambda, \lambda', \gamma$  and  $\gamma'$ , and can be made arbitrarily small at the expense of additional computing time (see Theorem 7 in Appendix C). The bias can also be completely removed by combining (3.18)-(3.19) with Metropolis-Hastings steps, as discussed in detail in [103]. However, doing this is not straightforward as calibrating the acceptance rate of a Metropolis-Hastings correction within a high-dimensional SAPG scheme can be very difficult (both automatically or manually). The reason for this is that the target density  $R_{\gamma, \lambda, \theta_n}$  changes at each iteration of the SAPG scheme, and every time the target density changes the parameters that control the acceptance rate of the Metropolis-Hastings steps need to be re-calibrated. If the acceptance rate is too low, too many samples might end up being discarded and this can significantly deteriorate the non-asymptotic convergence properties thus increasing the computing times [56]. For this reason we do not explore this any further.

### 3.1.5 Connections to the expectation-maximisation algorithm

Estimation problems of the form (2.8) can often be solved using the expectation-maximisation (EM) algorithm [51] or stochastic variations of it. More precisely, given an initial estimate of the parameter  $\theta_0 \in \Theta$ , the EM algorithm would solve (2.8) by iteratively repeating these two steps until convergence:

$$\begin{aligned} \text{E-step: } Q(\theta|\theta_n) &= \int_{\mathbb{R}^d} \log p(x, y|\theta) p(x|y, \theta_n) \, dx \\ \text{M-step: } \theta_{n+1} &= \underset{\theta \in \Theta}{\operatorname{argmax}} \, Q(\theta|\theta_n). \end{aligned} \tag{3.20}$$

The main difficulty in using the EM algorithm to compute (2.8) is that the E-step is intractable for most problems of interest. Several works have attempted to replace the E-step with different stochastic approximations [33, 50, 141]. In [6] they explain how SAPG schemes like the ones proposed in this thesis can be interpreted as an “approximate” stochastic EM algorithm where both the E and the M steps are only implemented approximately. To understand this connection, we can first rewrite

(3.20) replacing the M-step with a gradient step instead of the full maximisation [85]:

$$\begin{aligned} \text{E-step: } Q(\theta|\theta_n) &= \int_{\mathbb{R}^d} \log p(x, y|\theta) p(x|y, \theta_n) dx \\ \text{gradient-M-step: } \theta_{n+1} &= \theta_n + \delta \nabla_\theta Q(\theta_n|\theta_n) \\ &= \theta_n + \delta \int_{\mathbb{R}^d} \nabla_\theta \log p(x, y|\theta_n) p(x|y, \theta_n) dx. \end{aligned} \quad (3.21)$$

In our proposed method, we replace the expectation  $\int_{\mathbb{R}^d} \nabla_\theta \log p(x, y|\theta_n) p(x|y, \theta_n) dx$  in (3.21) with a Monte Carlo estimate  $\frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_{k+1}^n, y|\theta_n)$ , where  $X_{k+1}^n \sim R_{\gamma, \lambda, \theta_n}(X_k^n, \cdot)$  as detailed in (3.3). In this way, we can interpret Algorithm 1, Algorithm 2 and Algorithm 3 as generalised stochastic EM algorithms, where the expectation is approximated with a Monte Carlo estimate and the full maximisation is replaced by a gradient step.

This interpretation opens the door to an alternative variation of Algorithm 1 and Algorithm 2 where instead of using a gradient-M-step we can use a full M-step as the exact maximisation can be attained in closed form. For example, when  $m_n = 1$  the gradient estimate in Algorithm 1 is given by  $(\frac{d}{\alpha\theta} - g(X_{n+1}))$  which is zero for  $\theta = d/(\alpha g(X_{n+1}))$ . Therefore, we can use this to propose a modified scheme (see Algorithm 4 below) where the maximisation is performed exactly by setting  $\theta_{n+1} = \Pi_\Theta[d/(\alpha g(X_{n+1}))]$ .

---

**Algorithm 4** Variation of Algorithm 1 with exact maximisation step
 

---

- 1: **Input:** initial  $\{\theta_0, X_0\}$ ,  $(w_n)_{n \in \mathbb{N}}$ ,  $\Theta$ , kernel parameters  $\gamma, \lambda$ , iterations  $N$ .
  - 2: **for**  $n = 0$  to  $N - 1$  **do**
  - 3:     Sample  $X_{n+1} \sim R_{\gamma, \lambda, \theta_n}(X_n, \cdot)$ ,
  - 4:     Set  $\theta_{n+1} = \Pi_\Theta[d/(\alpha g(X_{n+1}))]$ .
  - 5: **end for**
  - 6: **Output:**  $\bar{\theta}_N$  computed with (3.5).
- 

This approach can also be extended to Algorithm 2 by applying the same idea to each separable component, thus leading to Algorithm 5 below. Our preliminary tests suggest that using the exact maximisation step can increase convergence speed (see Figure 4.12 in the hyperspectral unmixing experiment in Section 4.2 for a comparison between the evolution of the iterate  $\theta_n$  using Algorithm 2 and Algorithm 5).

However, this might not always be the case, especially when the limiting factor in the convergence speed is the sample correlation. Therefore more research is needed before we can draw stronger conclusions.

---

**Algorithm 5** Variation of Algorithm 2 with exact maximisation step

---

- 1: **Input:** initial  $\{\theta_0, X_0\}$ ,  $(\omega_n)_{n \in \mathbb{N}}$ ,  $\Theta$ , kernel parameters  $\gamma, \lambda$ , iterations  $N$ .
- 2: **for**  $n = 0$  to  $N - 1$  **do**
- 3:     Sample  $X_{n+1} \sim R_{\gamma, \lambda, \theta_n}(X_n, \cdot)$ ,
- 4:     **for**  $i = 1$  to  $d_\Theta$  **do**
- 5:         Set  $\theta_{n+1}^i = \Pi_{\Theta^i} \left[ |A_i| / \left( \alpha_i \tilde{g}_i(X_{n+1}[A_i]) \right) \right]$ .
- 6:     **end for**
- 7: **end for**
- 8: **Output:**  $\bar{\theta}_N$  computed with (3.5).

---

## 3.2 Example on synthetic data

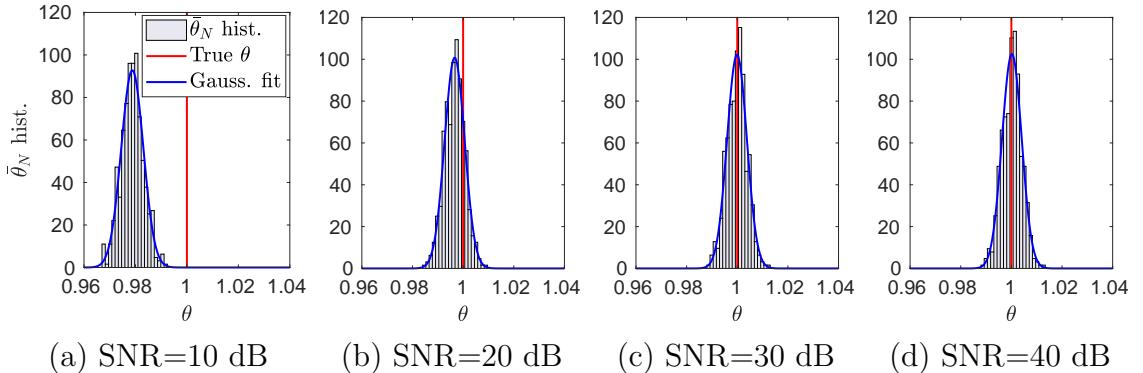
In this section we demonstrate the performance of the algorithm on a very simple image denoising problem, where we work with synthetic test images to have access to the true value of the regularisation parameter. The goal is to study the statistical behaviour of the algorithm as well as illustrate the role that each parameter plays in the algorithm. We also use this experiment to explore the robustness of the method towards mild likelihood misspecification (e.g., when there is a mismatch in the statistical properties of the noise).

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  $\Psi$  follow a Laplace distribution. That is the model (1.6) 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  $\sigma^2$  is chosen for every case such that the signal-to-noise-ratio (SNR) is 10 dB, 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 (3.18).

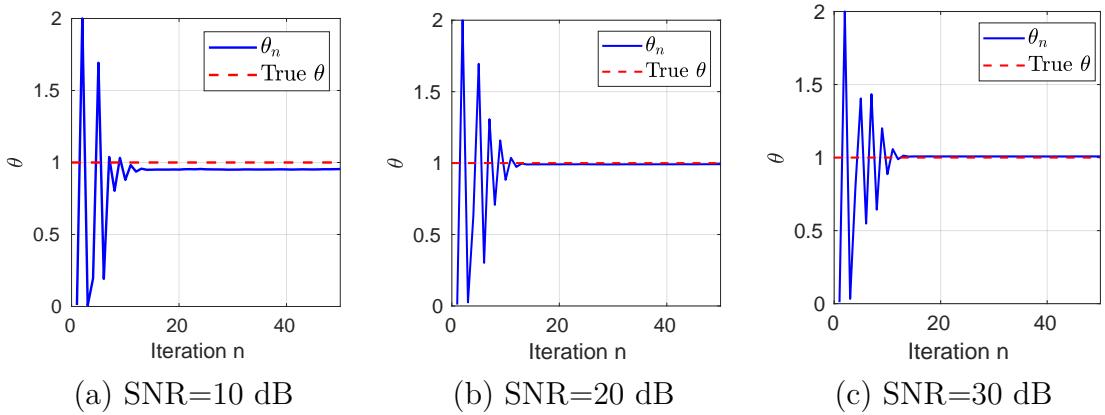
### 3.2.1 Estimation variance and bias

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

random  $x$ ; then, for every observation  $y$ , we estimate  $\bar{\theta}_N$ .



**Figure 3.1** – Denoising with synthesis- $\ell_1$  prior. Histograms of the estimated  $\bar{\theta}_N$  for 500 repetitions for different SNR values, using  $\lambda_{\max} = 2$ .



**Figure 3.2** – Denoising with synthesis- $\ell_1$  prior. Evolution of the iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different SNR values, using  $\lambda_{\max} = 2$ .

Figure 3.1 shows the histograms obtained from the 500 estimated  $\bar{\theta}_N$  values for each experiment (10 dB, 20 dB, 30 dB, and 40 dB). For completeness, we also present in Figure 3.2 one example of a generated sequence of iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different experiments, where we see that the algorithm converges in approximately 15 iterations. Observe that the estimation error is close to Gaussian and, for the higher SNR values ( $\text{SNR} \geq 20$  dB), the error is close to the true value of the regularisation parameter, as expected for a maximum likelihood estimator.

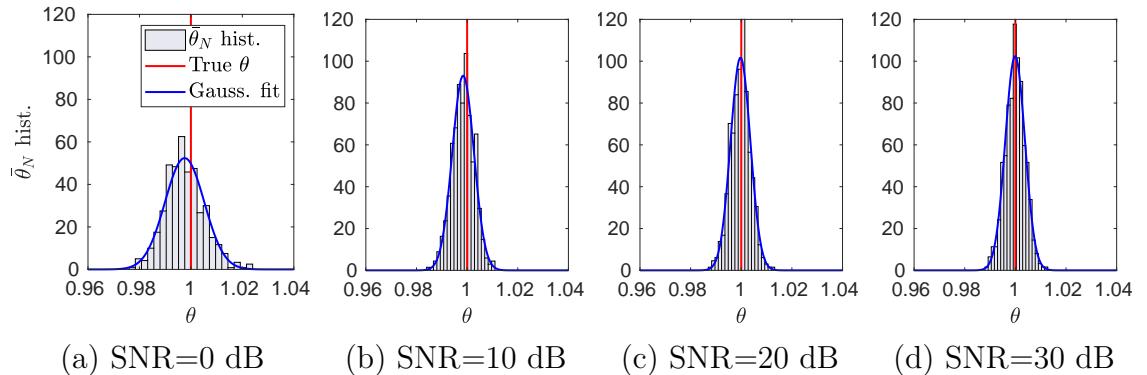
For the lowest SNR (10 dB), we see that the estimates exhibit a larger bias. The fact that the bias increases as the SNR decreases is a consequence of the way in which we set the algorithm parameters: we set  $\gamma$  and  $\lambda$  following the guidelines provided in Section 3.3.1, i.e. we set  $\gamma = 0.98(L_y + 1/\lambda)^{-1}$  and  $\lambda = \min(5L_y^{-1}, \lambda_{\max})$  with  $\lambda_{\max} = 2$ . In this way we use larger values of  $\gamma$  and  $\lambda$  for lower SNR values

and this improves the convergence speed. However, this rule of thumb is not always good enough for extremely low SNR values. One way of reducing the bias for low SNR setups is by using a smaller  $\lambda_{\max}$  at the expense of a slower convergence rate.

In Figure 3.3 we show the results obtained by repeating the same experiment from Figure 3.1 but using  $\lambda_{\max} = 0.0019604$  (which is the value previously used for the 30 dB experiment). Additionally, we also compute results for an extreme noise case (SNR=0 dB). We do not show the results for SNR=40 dB as they are practically identical to the ones shown in Figure 3.1.

As it may be seen, when using a smaller value of  $\lambda_{\max}$  the bias is effectively reduced: it significantly decreases for SNR=10 dB, and it almost disappears for SNR=20 dB. Even for the extreme noise case, the bias is relatively small (of the order of 0.2%) while the variance is much larger than for the other SNR values. If needed, the bias could be further reduced by decreasing  $\lambda_{\max}$  or by using a smaller  $\gamma$  at the expense of a slower convergence.

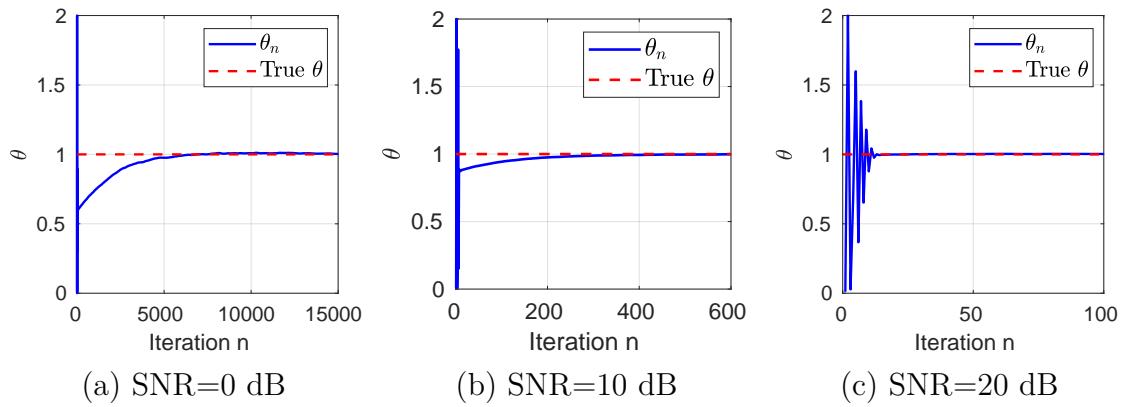
The reduction of the convergence speed can be appreciated in Figure 3.4, where we see that the algorithm converges in around 10000 iterations for SNR=0 dB, and 400 for SNR=10 dB.



**Figure 3.3 –** Denoising with synthesis- $\ell_1$  prior. Histograms of the estimated  $\bar{\theta}_N$  for 500 repetitions for different SNR values, using  $\lambda_{\max} = 0.0019604$ .

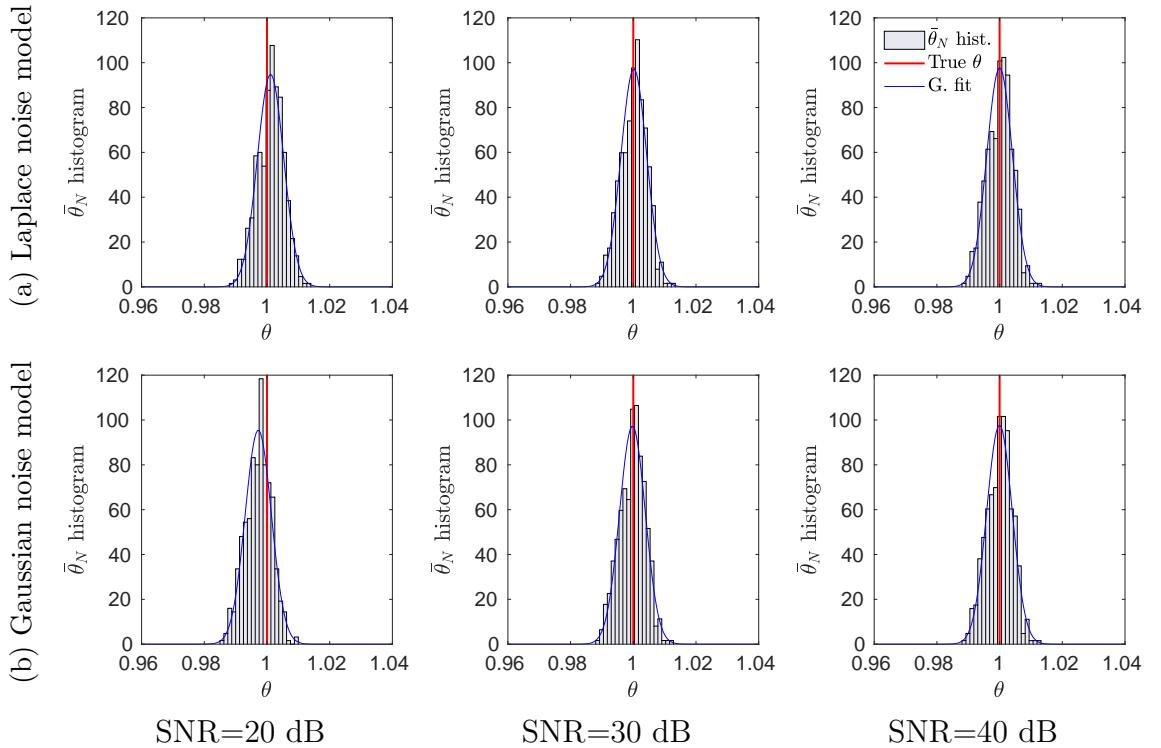
### 3.2.2 Laplace noise and likelihood misspecification

To explore the behaviour of the method with other noise distributions, we repeat the previous experiment using Laplace noise instead of Gaussian noise. Since the Laplace distribution involves a non-smooth  $\ell_1$  term, we adopt a proximal MCMC



**Figure 3.4** – Denoising with synthesis- $\ell_1$  prior. Evolution of the iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different SNR values, using  $\lambda_{\max} = 0.0019604$ .

approach and implement the algorithms using the gradient of its  $\lambda$ -Moreau-Yosida envelope.



**Figure 3.5** – Denoising with synthesis- $\ell_1$  prior with Laplace noise. Histograms of the estimated  $\bar{\theta}_N$  for 500 repetitions of the empirical Bayes algorithm using (a) correct Laplace noise model and (b) incorrect Gaussian noise model. Results for SNR of 20 dB, 30 dB and 40 dB.

The results are reported in Figure 3.5(a), where we observe that the method also performs well with Laplace noise.

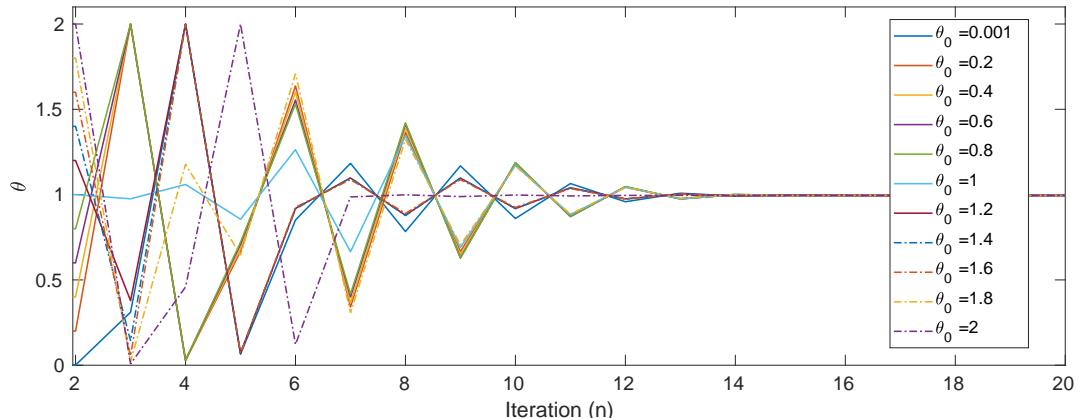
Lastly, to explore the robustness of the method towards mild misspecification

of the likelihood, we have also repeated the same experiment with Laplace noise but using an incorrectly specified Gaussian noise model (i.e. we generated the observation using Laplace noise but assumed the noise to be Gaussian in the model used to estimate  $\theta$ ). These results are reported in Figure 3.5 (b), where we see that the method is robust to mild likelihood misspecification, as the differences between using a correctly specified likelihood or an incorrectly Gaussian likelihood only become noticeable for the lowest SNR of 20 dB.

### 3.2.3 Role of the algorithm parameters

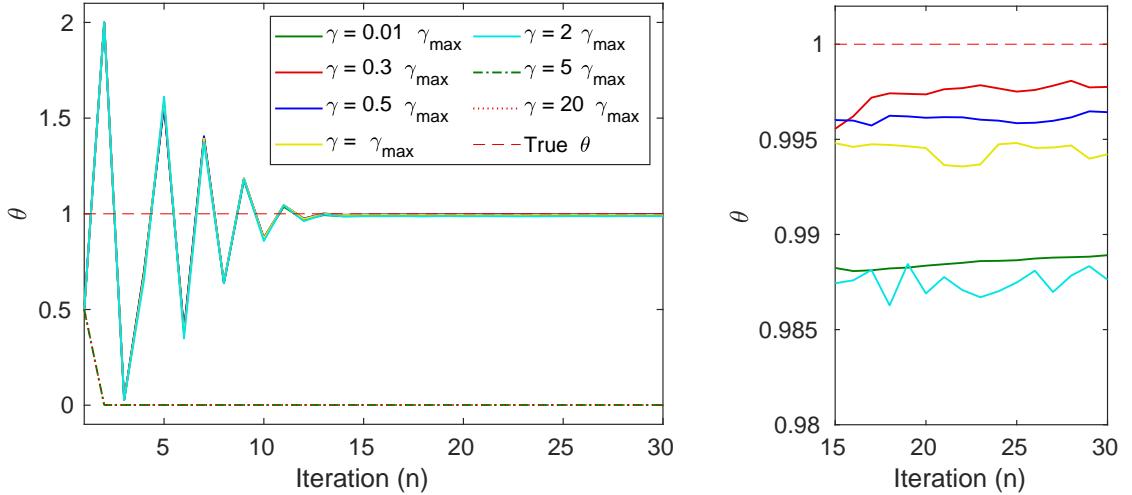
In this subsection we focus on individual executions of the algorithm and study the effect that different parameters have on the resulting behaviour of the proposed scheme.

**Initial  $\theta_0$**  From a theoretical point of view, the choice of the initial  $\theta_0 \in \Theta$  is asymptotically irrelevant. This is exemplified in Figure 3.6, where we see that regardless of the initialisation the algorithm converges to the same point. Nevertheless, in some cases a very bad initialisation can prevent the algorithm from converging, e.g., by introducing large numerical errors in the computation of proximal operators. (see Section 3.3.1).



**Figure 3.6** – Denoising with synthesis- $\ell_1$  prior. Evolution of the sequence of iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different initial values  $\theta_0$  for SNR=20 dB. All executions converge to the same value regardless of the initialisation.

**Step-size  $\gamma$  for the Markov kernels** As mentioned previously,  $\gamma$  controls the discretisation step-size of the continuous time Langevin diffusion. We know from [56] that  $\gamma$  should take values in the range  $(0, 2\gamma_{max})$  with  $\gamma_{max} = 1/(L_y + \lambda^{-1})$ , to guarantee the stability of the Euler-Maruyama discretisation.

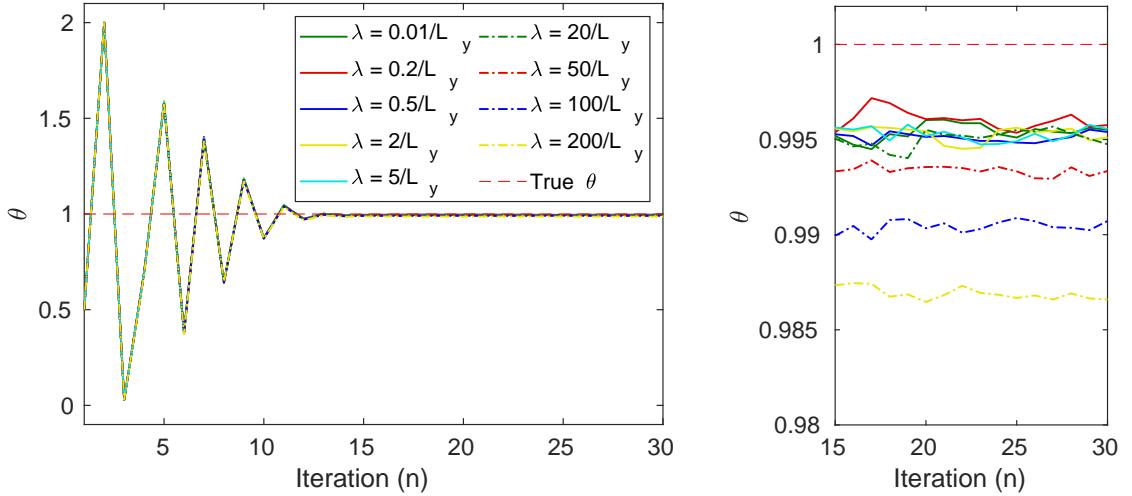


**Figure 3.7** – Denoising with synthesis- $\ell_1$  prior. Evolution of the sequence of iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different initial values  $\gamma$  for SNR=20 dB. All executions converge to the same value regardless of the initialisation.

In Figure 3.7 we illustrate the behaviour of the algorithm for different values of  $\gamma$  (we express the values relative to  $\gamma_{max}$ ). Notice that when  $\gamma > 2\gamma_{max}$  the algorithm does not converge and the iterates  $\theta_n$  saturate the bounds defined by  $\Theta$ . We observe that, in general, smaller values of  $\gamma$  lead to a smaller asymptotic bias, although for the smallest value of  $\gamma = 0.01\gamma_{max}$  the bias seems to be larger. This is only because we show the first 30 iterations of the algorithm and if more iterations were considered, this bias would slowly decrease too. Overall, we see that as long as the value of  $\gamma$  is within the admissible range, the algorithm converges, and a very small  $\gamma$  can deteriorate the convergence speed. This is because a very small  $\gamma$  defines a very small discretisation step-size in the MYULA sampler, which in turn leads to very correlated samples, thus slowing down the convergence of the SAPG scheme.

**Smoothing parameter  $\lambda$  for the Markov kernels** As explained in Section 3.1.4, this parameter controls the smoothing of the approximation  $g^\lambda(x)$  in the MYULA kernels and  $g^\lambda(x)$  can be brought arbitrarily close to  $g(x)$  by reducing  $\lambda$ . Since  $\lambda^{-1}$  limits the value of  $\gamma_{max}$ , it is usually good to select  $\lambda^{-1} \approx L_y$  (we discuss this in

more detail in Section 3.3.1). In Figure 3.8 we show the effect of using different values of  $\lambda$  and for clarity we express the values in terms of  $L_y$ . As expected, we observe that larger values of  $\lambda$  lead to a larger estimation bias.



**Figure 3.8** – Denoising with synthesis- $\ell_1$  prior. Evolution of the sequence of iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different values of  $\lambda$  for SNR=20 dB. The estimation bias increases with  $\lambda$ .

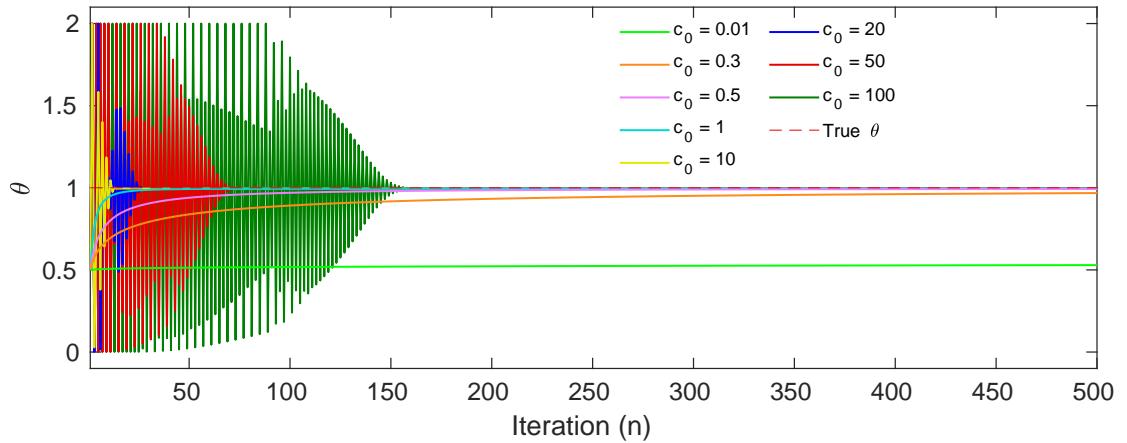
**Step-size  $\delta_n$  for the gradient step used to update  $\theta_n$**  We consider  $\delta_n = c_0 n^{-0.8}$ , which is a standard empirical choice in the literature<sup>4</sup> [17], and illustrate in Figure 3.9 the results obtained for different values of  $c_0$ . As it may be seen, when  $c_0$  is too large the iterate  $\theta_n$  oscillates for a long transient regime before converging. However, if  $c_0$  is too small, the convergence is significantly slowed.

### 3.3 Implementation guidelines

We now discuss suitable ranges and recommended values for the parameters of Algorithm 1, Algorithm 2 and Algorithm 3. Rather than optimal values for specific models, our recommendations seek to provide general rules that are simple and robust. We also discuss some other considerations related to the implementation and troubleshooting of the methods.

---

<sup>4</sup>We discuss the choice of  $\delta_n$  and  $c_0$  further in Section 3.3.1



**Figure 3.9** – Denoising with synthesis- $\ell_1$  prior. Evolution of the sequence of iterates  $(\theta_n)_{n \in \mathbb{N}}$  for different values of the scale  $c_0$  in  $\delta_n$  for SNR=20 dB.

### 3.3.1 Setting the algorithm parameters

**Selecting  $\gamma$**  In our theoretical convergence analysis (in Appendix C), Theorem 7 requires setting  $0 < \gamma < (L_y + 1/\lambda)^{-1}$ ; this is related to the numerical stability of the Markov chains and stems from the fact that  $L_y + 1/\lambda$  bounds the Lipschitz constant of  $\nabla f_y + (x - \text{prox}_{\theta^\top g}^\lambda(x)) / \lambda$ . Within this stability range,  $\gamma$  controls a trade-off between computational efficiency and accuracy, with larger values of  $\gamma$  leading to higher efficiency but also to a larger asymptotic bias. Given the dimensionality involved, and that in our experiments we did not observe any significant bias issues, we recommend using a large  $\gamma$ , e.g.,  $\gamma = 0.98(L_y + 1/\lambda)^{-1}$ .

**Selecting  $\lambda$**  This parameter controls the regularity of the smooth approximation of  $g$  within MYULA and hence another trade-off between bias and convergence speed [56]. We have empirically observed that, to prevent a significant bias, it is necessary to set  $\lambda \in (0, \lambda_{\max})$ , where for SNR  $\geq 20$  dB it is enough to set  $\lambda_{\max} = 2$ , and for very low SNR values  $\lambda_{\max}$  needs to be one or two orders of magnitude smaller (otherwise the estimation bias is too large, as shown in Section 3.2.1). Within this range, we prefer larger values of  $\lambda$  to improve convergence speed, at the expense of some bias. We recommend using  $\lambda = \min(L_y^{-1}, \lambda_{\max})$ , as setting  $\lambda \gg L_y^{-1}$  increases asymptotic bias without improving convergence speed because of the effect of  $L_y$  on  $\gamma$ .

**Selecting  $\gamma'$  and  $\lambda'$**  Since  $L_y$  does not affect the kernel  $\bar{R}_{\gamma', \lambda', \theta}$  targeting the prior, the stability range for  $\gamma'$  is  $0 < \gamma' < \lambda'$ . In our experiments we set  $\gamma' = 0.98\lambda'$ , but

any value of  $\gamma'$  that is close to but smaller than  $\lambda'$  will give similar results. We usually set  $\lambda' = \lambda$  to have the same level of smoothing in both chains, however one can also use  $\lambda' \gg \lambda$  if  $\bar{R}_{\gamma', \lambda', \theta}$  is much slower (i.e. the samples are more correlated) than  $R_{\gamma', \lambda', \theta}$ . It is important to highlight that the relative speed of both Markov kernels should be similar in order to improve the non-asymptotic convergence properties, especially when working with  $m_n = 1$ . If one kernel produces samples that are much more correlated than the ones coming from the other kernel, then some thinning is required (subsampling) in the slower kernel to help balance out the relative speeds. For further discussion of practical considerations regarding the relative speeds of the kernels see Section 3.3.5 and Section 3.3.7.

**Selecting  $(\delta_n, m_n)_{n \in \mathbb{N}}$**  For simplicity and computational efficiency, we recommend using a single ( $m_n = 1$ ) Monte Carlo sample per iteration. A single sample is sufficient to construct a convergent SAPG scheme (shown in Appendix C) provided that the sequence of gradient step-sizes  $(\delta_n)_{n \in \mathbb{N}}$  verifies

$$\sum_{n \in \mathbb{N}} \delta_n = +\infty \quad \text{and} \quad \sum_{n \in \mathbb{N}} \delta_n^2 < +\infty. \quad (3.22)$$

The first condition ensures that the gradient updates are large enough to asymptotically drive the iterates to a minimiser, and the second provides robustness w.r.t. the errors in the stochastic gradient estimates (see, e.g., [84]).

We recommend setting  $\delta_n = c_0 n^{-p}$  with  $p \in [0.6, 0.9]$ , and use  $\delta_n = c_0 n^{-0.8}$  in our experiments. This is a standard empirical choice in the literature [17] that verifies the requirements (3.22). For  $c_0$  we recommend, for the case where  $\theta$  is scalar, starting with  $c_0 = (\theta_0 d)^{-1}$  and then adjust if necessary. Although the choice of  $c_0$  is asymptotically irrelevant (see Figure 3.9), if the initial step-size is too large the iterate  $\theta_n$  will be bouncing on the limits of the interval for a long transient regime, whereas convergence will be slow if  $c_0$  is too small. For this reason, we recommend adjusting  $c_0$  so that the step-size is of the order of the projection interval  $\Theta$ . When  $\theta$  is not scalar, one can use different scales for each component of  $\theta$ . More details are provided in the Section 3.3.6.

**Selecting  $(\omega_n)_{n \in \mathbb{N}}$  and  $N$**  While it is possible to construct other estimates, we recommend using the average  $\bar{\theta}_N = \sum_{n=0}^{N-1} \omega_n \theta_n / \sum_{n=0}^{N-1} \omega_n$ , with  $(\omega_n)_{n \in \mathbb{N}}$  given by

$$\omega_n = \begin{cases} 0, & \text{if } n < N_0, \\ 1, & \text{if } N_0 \leq n \leq N_1, \\ \delta_n & \text{otherwise,} \end{cases} \quad (3.23)$$

where  $N_0, N_1 \in \mathbb{N}$ ,  $N_0 < N_1$ . This choice of  $(\omega_n)_{n \in \mathbb{N}}$ , defines three distinct phases: i) a burn-in phase where the first  $N_0$  iterations of the algorithm are discarded to reduce the non-asymptotic bias (this is particularly important when using a small number of iterations); ii) a uniform averaging phase  $N_0 \leq n \leq N_1$  where the smoothing effect associated with averaging improves convergence speed and reduces estimation variance; iii) a refinement phase where we use decreasing weights to improve the precision of the estimator (see Appendix C for accuracy guarantees).

We have empirically observed that imaging problems do not usually require highly accurate estimates of  $\theta$ . Therefore, in the interest of computational efficiency, in our experiments we omit the third phase and stop when  $N_1 = N$ . Moreover, rather than using the theoretical accuracy guarantees of Appendix C to set  $N$ , we monitor  $|\bar{\theta}_{N+1} - \bar{\theta}_N|/\bar{\theta}_N$  and stop the algorithms when  $|\bar{\theta}_{N+1} - \bar{\theta}_N|/\bar{\theta}_N < \tau$  for a prescribed tolerance  $\tau > 0$  (e.g.,  $\tau = 10^{-3}$ ).

**Selecting  $\Theta$**  When selecting the projection interval, the lower bound should be as small as necessary but not zero, as this may render the algorithm unstable (the gradient depends on  $\theta^{-1}$  and diverges as  $\theta_n \rightarrow 0$ ). If possible, use tight bounds to improve convergence speed. The choice of these bounds is empirical, depending on each particular problem.

**Selecting  $\theta_0$**  The choice of  $\theta_0 \in \Theta$  is theoretically asymptotically irrelevant (see, e.g., Figure 3.6). However, in some cases a very bad initialisation can prevent the algorithm from converging, e.g., by introducing large numerical errors in the computation of proximal operators. We observed this in the total generalised variation denoising experiment in Section 4.3: when using the extreme initialisation  $\theta_0^1 = \theta_0^2 = 100$  the algorithm did not converge due to numerical errors in the com-

putation of the proximal operator.

### 3.3.2 Other implementation considerations

We now provide some additional guidelines regarding the implementation and troubleshooting of the proposed methodology.

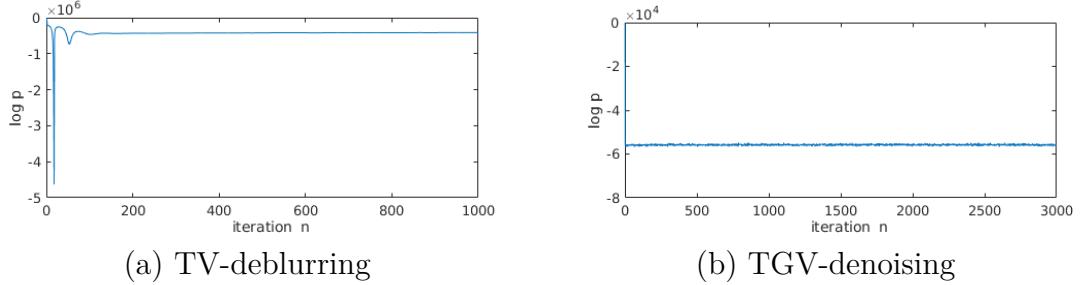
**Implementation in logarithmic scale** The proposed algorithms to estimate  $\theta$  often exhibit better numerical convergence properties when they are implemented in a logarithmic scale, which is a standard strategy for scale parameters [6]. Accordingly, we introduce the change of variables  $\eta = \log(\theta)$ , obtain an estimate  $\hat{\eta}$  by using one of the proposed algorithms to maximise the marginal likelihood  $p(y|\eta)$ , and then set  $\hat{\theta} = e^{\hat{\eta}}$ . This is equivalent to maximising  $p(y|\theta)$  because of the invariance to re-parametrisation property of the maximum likelihood estimator. This change of variables requires a minor modification in the computation of the gradients, which have to be multiplied by  $e^\eta$  to satisfy the chain rule. For example, step 9 in Algorithm 1 becomes  $\eta_{n+1} = \Pi_{\Theta^\eta} \left[ \eta_n + e^{\eta_n} \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \left\{ \frac{de^{-\eta_n}}{\alpha} - g(X_k^n) \right\} \right]$ , where  $\Theta^\eta = \{\log(\theta) : \theta \in \Theta\}$  denotes the range of admissible values of  $\eta$  taking the logarithm component-wise. Similarly, step 11 of Algorithm 3 becomes  $\eta_{n+1} = \Pi_{\Theta^\eta} \left[ \eta_n + e^{\eta_n} \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \{g(\bar{X}_k^n) - g(X_k^n)\} \right]$ .

**Initialisation of the Markov kernels** We strongly recommend warm-starting the Markov chains by running  $T_0 \in \mathbb{N}$  iterations with fixed  $\theta = \theta_0$  before starting to update the value of  $\theta$ ; an appropriate value for  $T_0$  can be easily determined by monitoring the log-probability and running warm-up iterations until it becomes stable, see Section 3.3.3 for details.

### 3.3.3 Testing the MCMC sampler

Before trying to adjust the value of  $\theta \in \Theta$  with the algorithm, we strongly recommend starting by testing the MCMC sampler with a fixed value of  $\theta$ . A simple way to see whether the Markov chain is working as expected, is to plot something proportional to the value of the log-probability of the samples; we typically plot  $-f_y(X_k^n) - \theta^T g(X_k^n)$  for the posterior distribution and  $-\theta^T g(X_k^n)$  for the prior.

As mentioned in Section 1.2, there is a useful concentration phenomenon studied in [16, Theorem 1.2] which implies that for high-dimensional log-concave densities  $\pi$ , a Markov chain targeting  $\pi$  eventually starts generating samples  $X_k^n$  for which  $\log \pi(X_k^n)$  is approximately constant (and close to the entropy). Therefore, if the MCMC sampling is successful the log-probability stabilises after some iterations and remains more or less constant.



**Figure 3.10** – Evolution of  $(\log p(X_1^n | y, \theta))_{n \in \mathbb{N}}$  with  $(X_1^n)_{n \in \mathbb{N}}$  sampled using MYULA and targeting  $p(\cdot | y, \theta)$ . Results for (a) TV-deblurring with SNR = 40dB and (b) TGV-denoising with SNR = 8dB.

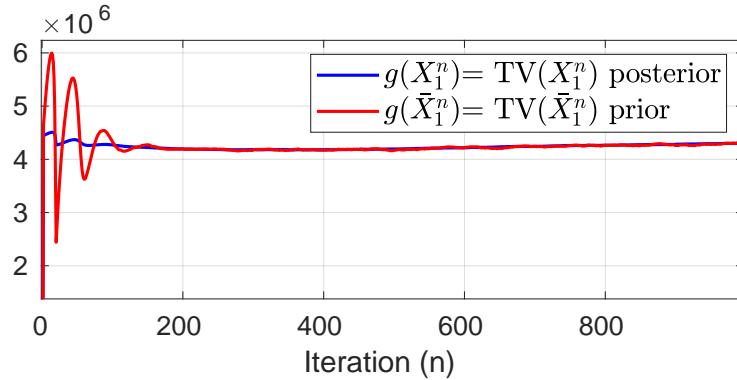
Conversely, if plots show that the chain is divergent or very unstable, then there might be a problem with the sampler. A common cause for divergence is setting a discretisation step-size that is too large. We would advise not to proceed with the estimation of  $\theta$  until the sampler shows a stable behaviour similar to the ones shown on Figure 3.10.

### 3.3.4 Monitoring convergence in Algorithm 1, Algorithm 2 and Algorithm 3

**Lack of convergence due to bound saturation** If one observes that the iterate  $\theta_n$  saturates the limits of the projection interval  $\Theta$ , one should first verify that the Markov kernels are working properly (see recommendations in 3.3.3). If they are, then the problem might be that the solution lies outside  $\Theta$ . If  $\theta$  is multivariate and only some components are saturating the bounds, then check the scale and projection bounds for those specific components.

**Verifying proper convergence** As the algorithm converges, the iterates  $\theta_n$  get closer to a maximiser of  $p(y|\theta)$  and the gradient estimates  $\Delta_{m_n, \theta_n}$  vanish in expectation. Hence, the residual  $\|\Delta_{m_n, \theta_n}\|$  should become small (on average) as  $n$

increases, i.e.,  $g(X_k^n)$  will become close to  $d/(\alpha\theta_n)$  in Algorithm 1, or close to  $g(\bar{X}_k^n)$  in Algorithm 3<sup>5</sup>. It is therefore useful to plot the traces of  $(g(X_{k=k_0}^n))_{n \in \mathbb{N}}$  together with  $(g(\bar{X}_{k=k_0}^n))_{n \in \mathbb{N}}$  or  $(d/(\alpha\theta_n))_{n \in \mathbb{N}}$  as appropriate, to check that the algorithm is converging. The trace can be plotted for a fixed value of  $k = k_0$  as this is enough to monitor the convergence. This is illustrated for Algorithm 3 in Figure 3.11 below, where we observe how these terms become closer as the number of iterations increases.



**Figure 3.11** – Evolution of the iterates  $(g(X_1^n))_{n \in \mathbb{N}}$  and  $(g(\bar{X}_1^n))_{n \in \mathbb{N}}$  for the proposed method in a deblurring experiment with a TV prior and SNR = 40dB.

If  $\|\Delta_{m_n, \theta_n}\|$  does not vanish as  $n$  increases this could indicate a problem with the choice of  $\delta_n$  or that the two MCMC kernels have very different speed (See Section 3.3.5).

### 3.3.5 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 deteriorate the convergence properties of the SAPG algorithm.

This kind of imbalance can be detected by plotting the sample autocorrelation for each chain using  $g$  as a summary statistic. If the autocorrelation plots decay at significantly different rates, it is necessary to reduce the correlation within the slower chain by either introducing some thinning (which essentially amounts to concatenating several iterations of the kernel to improve its convergence speed) or by increasing the step-size  $\gamma$  (see Section 3.3.7).

---

<sup>5</sup>For Algorithm 2 use a component-wise comparison between  $\frac{|\mathbf{A}_i|}{\alpha_i \theta_n^2}$  and  $\tilde{g}_i\left(X_{k[\mathbf{A}_i]}^n\right)$  for every  $i \in \{1, \dots, d_\Theta\}$

Notice that it is especially important to pay attention to the relative speeds of the kernels if one of the MYULA kernels was replaced by, for instance, a perfect sampling method.

### 3.3.6 Working with multivariate $\theta$

When  $\theta$  is multivariate each component of the solution might have a different order of magnitude. In this case, we recommend using different step-size scales for each component of  $\theta$ . For example, we can compute  $\theta_n = \Pi_\Theta [\theta_n + D \delta_n \Delta_{m_n, \theta_n}]$ , where  $D \in \mathbb{R}^{d_\Theta \times d_\Theta}$  is a diagonal matrix, and each element of the diagonal scales one component of  $\theta$ . It is also helpful to remember that one can run the algorithm with some components of  $\theta$  fixed. This allows isolating components and verifying convergence for subsets of  $\theta$ .

The possibility of using second order methods for the update of  $\theta$  is not explored in this these but it is a perspective for future work, especially for problems with multivariate  $\theta$  with strong dependencies between the different components.

### 3.3.7 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  $\gamma$  must 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  $\lambda'$  of this particular kernel, in order to be able to increase the value of the discretisation step-size  $\gamma'$ .

In more general cases where the limiting factor for  $\gamma$  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 4.2) to reduce gradient anisotropy and improve the condition number of the problem. This is a standard practice which consists of re-scaling the problem to improve its condition number,

i.e. to make the problem more “isotropic”. Let  $P = (A^\top A)^{-1}$  and let  $P^{1/2}$  be the matrix square root of  $P$ , then the MYULA kernel can be written as

$$x^{(t+1)} = x^{(t)} - \gamma P \nabla f_y(x^{(t)}) - \gamma P \frac{\left( x^{(t)} - \text{prox}_{\theta^\top g}^\lambda(x^{(t)}) \right)}{\lambda} + P^{1/2} \sqrt{2\gamma_m} z^{(t+1)}, \quad (3.24)$$

If the matrix  $P$  is not available, it can be learnt online using the technique from [5].

For cases where the problem is severely ill-posed an alternative strategy is to use the tamed unadjusted Langevin algorithm proposed in [19]. This requires replacing the gradient  $\nabla f_y$  by one of the two possible “tamed” versions. Either the gradient is replaced by its globally tamed version

$$\frac{\nabla f_y(x)}{1 + \gamma \|\nabla f_y(x)\|_2}, \quad (3.25)$$

or it is replaced by its component-tamed version

$$\left( \frac{\partial_i f_y(x)}{1 + \gamma \|\partial_i f_y(x)\|_2} \right)_{i \in \{1, \dots, n\}}, \quad (3.26)$$

which tends to give better results. It is worth noting that this approach may introduce a small additional bias, as explained in [19].

**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.

### 3.3.8 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  $\gamma$  and the smoothing parameter  $\lambda$ . However, based on what we have observed in practice, the limiting factor tends to be  $\lambda$ . If there is a bias issue, we recommend trying to reduce  $\lambda$  to obtain a better approximation of the target distribution, at the expense

of some deterioration in convergence speed. 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 (subsampling) in the slower chain, as suggested in Section 3.3.5.

Alternatively, if the bias cannot be removed without a severe deterioration of the computing times, other approaches could be considered to approximate the expectations in (3.12). For instance, one could first use a variational approximation [77] of each distribution and sample from these surrogate distributions instead. We do not explore this in this thesis, but it is an interesting perspective for future work.

# Chapter 4

## Numerical experiments on imaging problems

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. In Section 4.1, we demonstrate the method by estimating a scalar-valued regularisation parameter in a non-blind (and non-myopic) image deconvolution model with different kinds of prior distributions, such as total variation and  $\ell_1$ -wavelet priors. This allows comparing our method to some state-of-the-art approaches that are limited to scalar-valued regularisation parameters. We also use one of these experiments to explain how to address problems in which the noise variance is unknown by jointly estimating  $\theta$  and the variance of the noise by marginal MLE.

This is then followed by two challenging problems involving multivariate regularisation parameters. In particular, in Section 4.2 we apply our method to a sparse hyperspectral unmixing problem combining an  $\ell_1$  and a total variation regularisation, and where we report comparisons with the hierarchical Bayesian approach of [105]. Lastly, in Section 4.3 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 (3.5), and then calculate a MAP estimator using the empirical Bayesian posterior  $x \mapsto p(x|y, \bar{\theta}_N)$  by convex opti-

misation (solver details are provided in each experiment)<sup>1</sup>. In all experiments,  $\theta$  was estimated on a logarithmic scale by using the change of variables discussed in Section 3.3.2. All experiments were conducted on an Intel i9-8950HK@2.90GHz running MATLAB R2018a.

## 4.1 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- $\ell_1$  prior. For comparison, we also report the results obtained with SUGAR [49] (only when using a TV prior), joint MAP estimation [105], discrepancy principle [58, 101], and by using the oracle value  $\theta_\dagger$  that minimises the estimation mean squared error (MSE), *i.e.*

$$\theta_\dagger = \arg \min_{\theta \in \Theta} \left\{ \left\| x^0 - \arg \max_{x \in \mathbb{R}^d} p(x|y, \theta) \right\|_2 \right\}, \quad (4.1)$$

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 B.

In non-blind image deblurring, the aim is to recover an unknown image  $x \in \mathbb{R}^d$  from a blurred and noisy observation  $y = Ax + w$ , where  $A \in \mathbb{R}^d \times \mathbb{R}^d$  is a blur matrix, and  $w$  is a  $d$ -dimensional Gaussian random variable with zero mean and covariance matrix  $\sigma I_d$  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  $9 \times 9$  pixels, and  $\sigma^2$  is chosen such that the blurred signal-to-noise-ratio (SNR) is 20 dB, 30 dB, or 40 dB. We define the blurred SNR (in dB) as

$$SNR = 10 \log_{10} \frac{\|Ax - \bar{A}x\|_2^2}{d \sigma^2}, \quad (4.2)$$

---

<sup>1</sup>We compute the MAP estimator as this is a standard practice for the experiments we consider and many of the convex optimisation solvers we use have been specifically designed for MAP estimation. However other estimators such as the MMSE could also be considered.

where  $\bar{A}x$  is obtained taking the average value of all pixels in  $Ax$ . We perform all experiments on ten standard test images (`barbara`, `boat`, `bridge`, `flintstones`, `goldhill`, `lake`, `lena`, `mandrill` and `wheel`).

For each image, noise level, and  $\theta$  selection method, we first obtain an estimate for  $\theta$  and then use it to compute the MAP estimator  $\hat{x}_{\text{MAP}}$  (given by (1.7)). In the case of the joint MAP method [105], we carry out joint MAP estimation of  $\theta$  and  $\hat{x}_{\text{MAP}}$ . We compute the MAP estimator by using a highly efficient proximal convex optimisation algorithm, SALSA [2], 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.

#### 4.1.1 Deconvolution with total variation prior

In this experiment we use model (1.6) where for any  $x \in \mathbb{R}^d$  we have  $f_y(x) = \frac{\|y - Ax\|_2^2}{2\sigma^2}$ ,  $g(x) = \text{TV}(x)$ , and follow the previously explained procedure. Here  $\text{TV}(x)$  is the isotropic total variation pseudo-norm given by  $\text{TV}(x) = \sum_i \sqrt{(\Delta_i^h x)^2 + (\Delta_i^v x)^2}$  where  $\Delta_i^v$  and  $\Delta_i^h$  denote horizontal and vertical first-order local difference operators. To compute  $\bar{\theta}_N$  we use Algorithm 1. The prior associated with the total variation pseudonorm is not proper [118, Section 1.5] (i.e.,  $\int p(x|\theta)dx = \infty$ ). This is because  $\text{TV}(x)$  does not depend on the average value of the pixels and only depends on the differences between them. This means that if we add a constant value to all pixels the total variation pseudonorm remains the same and therefore the effective dimension is  $d - 1$ . We evaluated the proximal operator of  $\text{TV}(x)$  using the primal-dual algorithm from [35] with 25 iterations.

The algorithm parameters are chosen following the recommendations provided in Section 3.3.1; we consider 300 warm-up iterations and set  $\theta_0 = 0.01$ ,  $X_0^0 = y$ ,  $m_n = 1$ ,  $\delta_n = 10 \times n^{-0.8}/d$  for any  $n \in \mathbb{N}^*$ , we set  $\lambda = \min(5L_y^{-1}, \lambda_{\max})$  with  $\lambda_{\max} = 2$  and  $L_y = (0.99/\sigma)^2$ , and  $\gamma = 0.98 \times (L_y + 1/\lambda)^{-1}$ . As suggested in Section 3.3.1, we set  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 25$  burn-in iterations and compute  $\bar{\theta}_N$  using (3.5).

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` =  $\bar{\theta}_N/10$ .



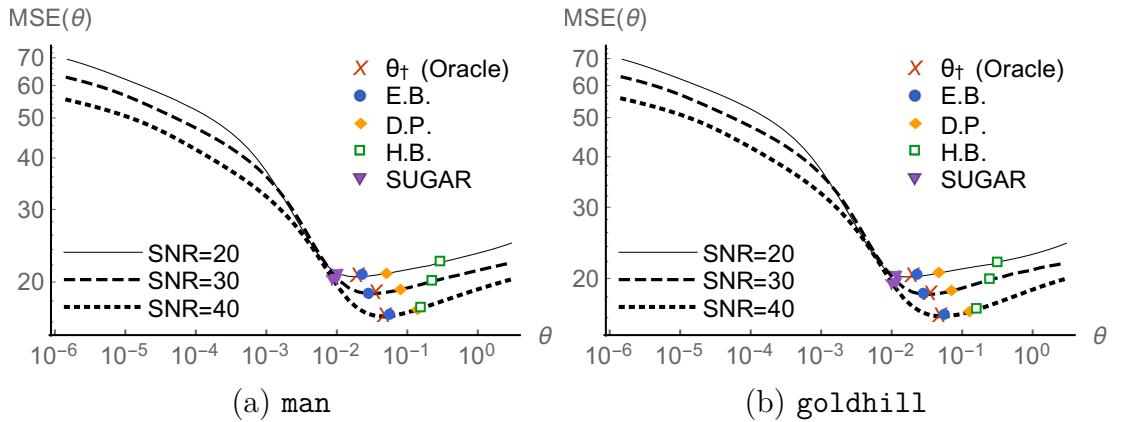
**Figure 4.1** – Deblurring with TV prior for `man` and `goldhill` test images: (a) blurred and noisy (SNR=30 dB) observation  $y$ , (b) MAP estimator obtained using  $\bar{\theta}_N$  computed with empirical Bayes.

For illustration, Figure 4.1 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 4.2 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 4.3 and Figure 4.4 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.

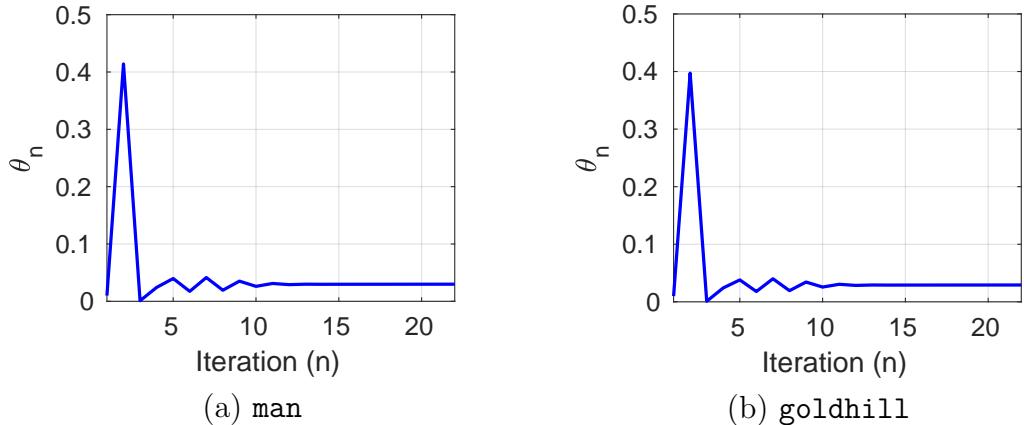
Observe in Figure 4.3 that the proposed empirical Bayesian algorithm yields



**Figure 4.2** – 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.



**Figure 4.3** – Deblurring with TV prior. Mean squared error (MSE) obtained for (a) `man` and (b) `goldhill` for different values of  $\theta$ . We compare the values obtained with empirical Bayes, discrepancy principle, hierarchical Bayes, SUGAR, and the optimal value  $\theta_{\dagger}$  that minimises the MSE.



**Figure 4.4** – 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).

close-to-optimal results, for both 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  $\theta$  (this can also be observed in the recovered image in Figure 4.2 (f), where the MAP estimate presents some ringing artefacts due to high-frequency noise amplification); this is in agreement with the results reported in [93].

Table 4.1 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 (obtained by using the oracle value  $\theta_\dagger$  as defined in (4.1)), 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 [105] 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  $\theta$  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 (see Appendix B for details about the comparison with SUGAR).

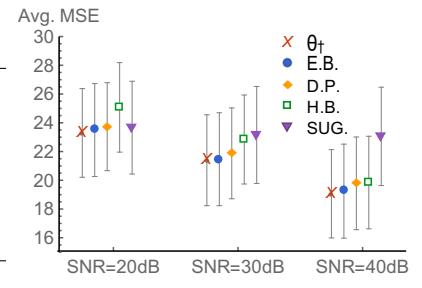
We emphasise at this point that the exact computing times of each algorithm depend on the specific stopping criteria and implementation details, so rather than claiming that one method is faster than the others, what we wish to illustrate is that the computing times are all within the same order of magnitude, 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 times would have increased with almost negligible changes in the MSE. Also note that we compute the optimal  $\theta$  for the discrepancy principle method by continuation, but one could also use a different proximal splitting strategy (see [42] for instance).

#### 4.1.2 Deconvolution with Total Variation prior and unknown noise variance

In this section we consider the same experiment as in Section 4.1.1, but we now suppose that the noise variance is unknown and explain how to modify our method-

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	MSE	Time	MSE	Time	MSE	Time
$\theta_\dagger$	23.29		21.39		19.06	
EB	<b>23.50</b>	0.84	<b>21.45</b>	0.85	<b>19.24</b>	0.85
DP	23.73	0.70	21.87	1.52	19.78	3.92
HB	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 4.1** – Deblurring with TV prior. (a) Table with average mean squared error (MSE) obtained for ten images with different algorithms (results with lowest MSE highlighted in bold). Average execution times expressed in minutes. In (b) we summarise the content of the table and show the standard deviation with error bars.

ology to estimate this quantity jointly with  $\theta$  by marginal MLE. This is beyond the scope of the theoretical results we present in Appendix C. However, we believe that the theory could be generalised to provide some (albeit weaker) guarantees for this case and other blind and semi-blind problems, and this is an important perspective for future work. Alternatively, the noise variance could also be pre-estimated with some other method such as the mean absolute derivative rule proposed in [52].

More precisely, we can use the proposed scheme to compute

$$(\theta_*, \sigma_*^2) \in \underset{\theta \in \Theta, \sigma^2 \in [\sigma_{min}^2, \sigma_{max}^2]}{\operatorname{argmax}} p(y|\theta, \sigma^2), \quad (4.3)$$

where  $0 < \sigma_{min}^2 < \sigma_{max}^2 < \infty$  define a minimum and maximum admissible variance values. To obtain an estimate of  $\frac{d}{d\sigma^2} \log p(y|\theta, \sigma^2)$  in Algorithm 1 we differentiate  $\log p(x, y|\theta, \sigma^2)$  w.r.t.  $\sigma^2$  and obtain

$$\frac{d}{d\sigma^2} \log p(x, y|\theta, \sigma^2) = \frac{\|y - Ax\|_2^2}{2(\sigma^2)^2} - \frac{d}{2\sigma^2}. \quad (4.4)$$

We summarise the resulting scheme for jointly estimating  $\theta$  and  $\sigma^2$  in Algorithm 6 below.

One of the complications that stems from working with an unknown noise variance is that the Lipschitz constant  $L_y$  is unknown. This is a problem because  $L_y$  affects the maximum step-size  $\gamma$  that we can use in the Markov kernels while ensuring convergence;  $L_y$  is usually also used to set  $\lambda$ . To overcome this, we propose to initialise the algorithm by assuming the worst-case scenario, i.e.  $\sigma^2 = \sigma_{min}^2$ , which

---

**Algorithm 6** SAPG algorithm - Scalar  $\theta$  and unknown noise variance  $\sigma^2$ 


---

```

1: Input: initial  $\{\theta_0, X_0^0\}$ ,  $(\delta_n, \delta'_n, \omega_n, m_n)_{n \in \mathbb{N}}$ ,  $\Theta$ , kernel parameters  $\gamma, \lambda$ , iterations  $N$ .
2: for  $n = 0$  to  $N - 1$  do
3:   if  $n > 0$  then
4:     Set  $X_0^n = X_{m_{n-1}}^{n-1}$ ,
5:   end if
6:   for  $k = 0$  to  $m_n - 1$  do
7:     Sample  $X_{k+1}^n \sim R_{\gamma, \lambda, \theta_n, \sigma_n^2}(X_k^n, \cdot)$ ,
8:   end for
9:   Set  $\theta_{n+1} = \Pi_\Theta \left[ \theta_n + \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \left\{ \frac{d}{\alpha \theta_n} - g(X_k^n) \right\} \right]$ .
10:  Set  $\sigma_{n+1}^2 = \Pi_{[\sigma_{min}^2, \sigma_{max}^2]} \left[ \sigma_n^2 + \frac{\delta'_{n+1}}{m_n} \sum_{k=1}^{m_n} \left\{ \|y - AX_k^n\|_2^2 / 2(\sigma_n^2)^2 - d / (2\sigma_n^2) \right\} \right]$ .
11: end for
12: Output:  $\bar{\theta}_N = \sum_{n=0}^{N-1} \omega_n \theta_n / \sum_{n=0}^{N-1} \omega_n$  and  $\bar{\sigma}_N^2 = \sum_{n=0}^{N-1} \omega_n \sigma_n^2 / \sum_{n=0}^{N-1} \omega_n$ .

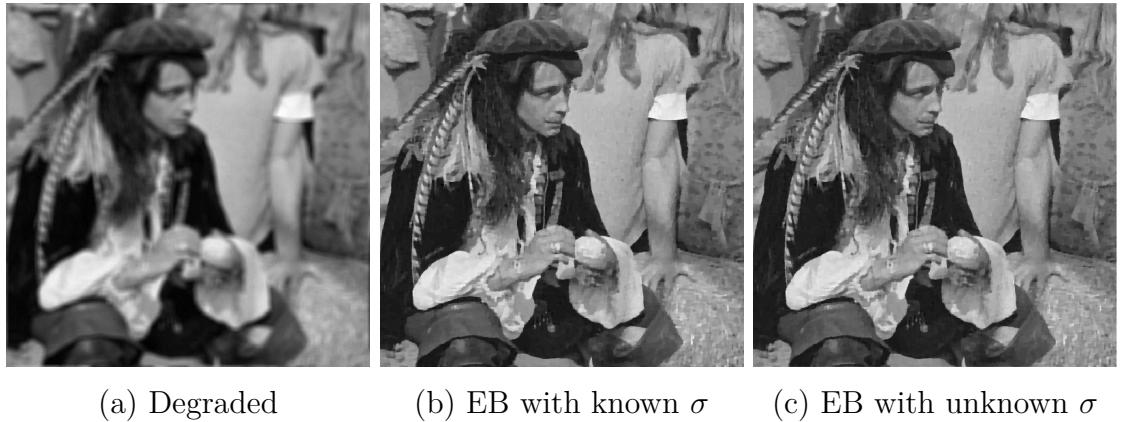
```

---

will lead to the largest  $\hat{L}_y = (0.99/\sigma_{min})^2$ , and in turn lead to the smallest possible step-size  $\gamma$  and a small  $\lambda$ . Since this value is usually very conservative, one can run some iterations of the algorithm until the value of  $\sigma_n^2$  begins to stabilise, then refine  $\hat{L}_y$  to update the algorithm parameters  $\gamma$  and  $\lambda$ , and continue iterations with those updated values. Here we adopt this approach and run the algorithm in three stages, where we update  $\gamma$  and  $\lambda$  at the end of each stage by using the estimates of  $\bar{\sigma}_N^2$  available at that point to refine  $\hat{L}_y$ . In accordance with the guidelines provided in Section 3.3.1, we set  $\lambda = \min(5\hat{L}_y^{-1}, \lambda_{\max})$  with  $\lambda_{\max} = 2$  and  $\gamma = 0.98 \times (\hat{L}_y + 1/\lambda)^{-1}$ . We have set  $\sigma_{min}^2$  and  $\sigma_{max}^2$  by assuming prior knowledge that the SNR is between 15 dB and 45 dB, but other values could be used without significantly impacting results. In each stage we use 300 warm-up iterations, set  $\theta_0 = 0.01$ ,  $\sigma_0^2 = (\sigma_{min}^2 + \sigma_{max}^2)/2$ ,  $X_0^0 = y$ ,  $m_n = 1$ ,  $\delta_n = 10 \times n^{-0.8}/d$ , and  $\delta'_n = 10 \times n^{-0.8}/d$  for any  $n \in \mathbb{N}^*$ . At each stage, we use the same stopping criteria as in Section 4.1.1, with a tolerance of  $10^{-3}$  for both  $\theta_n$  and  $\sigma_n$  (the algorithm progresses to the next stage (or is stopped) when both iterates meet the criteria).

For illustration, Figure 4.5 shows the results obtained with Algorithm 6 for the `man` test image. For comparison, we also show the results of Section 4.1.1 obtained by using the true value of  $\sigma$ . The displayed images correspond to the 30 dB SNR setup. Observe there is very little difference between the recovered image using the true value of  $\sigma^2$  and the marginal MLE estimate  $\bar{\sigma}_N^2$  obtained with Algorithm 6.

Table 4.2 presents a detailed comparison of the results obtained with Algo-



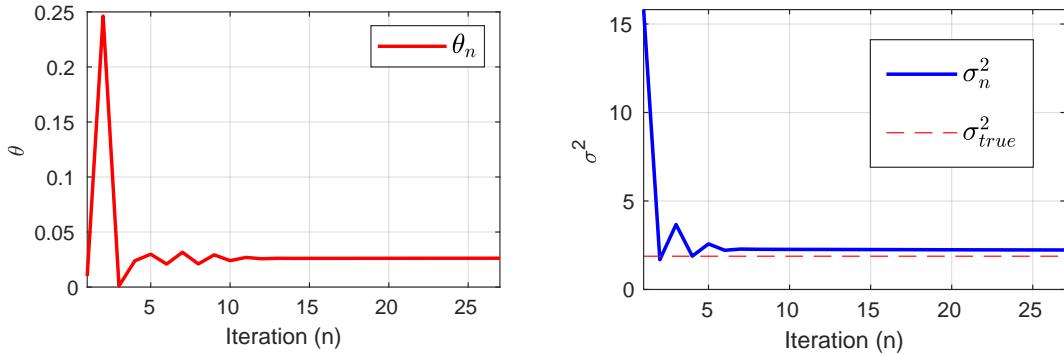
**Figure 4.5** – Deblurring with TV prior for `man`: (a) blurred and noisy (SNR=30 dB) observation  $y$ , (b-c) MAP estimator with  $\bar{\theta}_N$  computed with empirical Bayes using (b) true and (c) estimated  $\sigma$ .

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	MSE	Time (min)	MSE	Time (min)	MSE	Time (min)
$\theta_\dagger$	23.29		21.39		19.06	
EB with known $\sigma$	23.50	0.84	21.45	0.85	19.24	0.85
EB with unknown $\sigma$	23.53	1.02	21.52	1.35	19.27	1.77

**Table 4.2** – Deblurring with TV prior and unknown  $\sigma$ . Table with average mean squared error obtained for ten images for the experiment where  $\sigma$  is estimated jointly with  $\theta$ . For reference we also include the results obtained with empirical Bayes when  $\sigma$  is known and using the oracle value  $\theta_\dagger$  that minimises the MSE.

rithm 6. Again, observe that the quality of the restored images obtained with the marginal MLE estimate  $\bar{\sigma}_N^2$  is comparable to that of the images obtained with the true value of  $\sigma^2$ , with a moderate overhead in the computing times when the three-stage approach is used. This additional computing time is due to the fact that since  $L_y$  is not known, we start the algorithm with a very small step-size  $\gamma$  that, as explained, we iteratively increase. In general, using a smaller step-size leads to longer computing times (slower convergence). This is also reflected in the fact that in this and most other experiments, the computing times increase with SNR. The reason for this is that a larger noise variance  $\sigma$  leads to a smaller  $L_y$  which, in turn, tends to result in a larger step-size  $\gamma$  (unless  $\lambda$  is the limiting factor).

We conclude by presenting in Figure 4.6 the evolution of the iterates  $(\theta_n)_{n \in \mathbb{N}}$  and  $(\sigma_n^2)_{n \in \mathbb{N}}$  for the last stage of the algorithm (the first two stages are discarded). Observe that the algorithm converges very quickly, similarly to the case when  $\sigma^2$  is known.



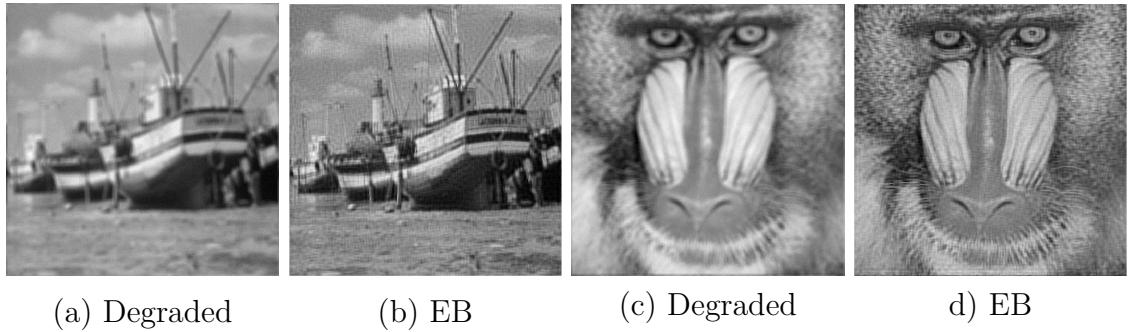
**Figure 4.6** – Deblurring with TV prior and unknown noise variance  $\sigma^2$ . Evolution of the sequence of iterates  $(\theta_n)_{n \in \mathbb{N}}$  and  $(\sigma_n^2)_{n \in \mathbb{N}}$  during the last stage of the proposed method for the `man` test image (SNR=30 dB).

#### 4.1.3 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 a redundant 4-level Haar wavelet representation  $\Psi$ , with dimension  $d = 10 \times d_y = 10 \times 512 \times 512$  coefficients. We assume a Laplace prior (i.e.,  $p(x|\theta) \propto e^{-\theta\|x\|_1}$ ) on the elements of  $x$  with unknown parameter  $\theta$ . Accordingly, the posterior is of the form (1.6) with  $f_y(x) = \|y - Ax\|_2^2 / (2\sigma^2)$ ,  $g(x) = \|x\|_1$ . To obtain solutions we map  $x$  to pixel domain by computing  $\Psi^\top x$ .

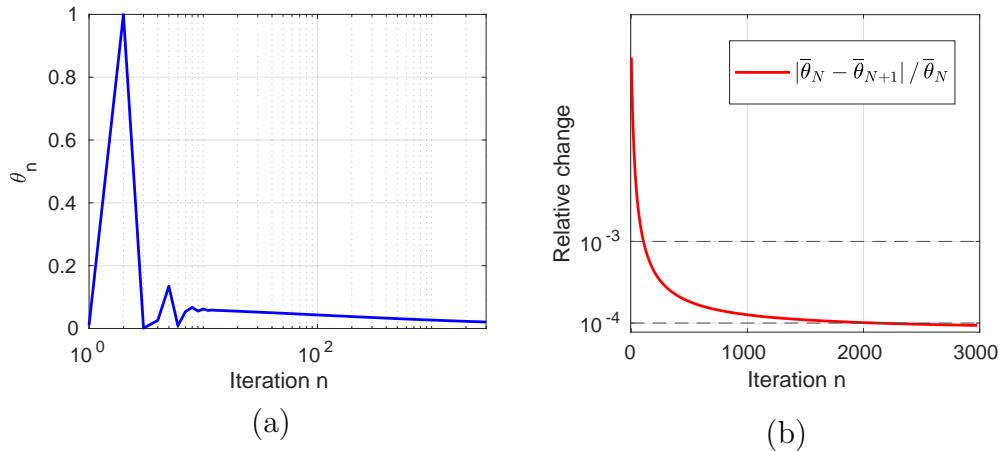
To compute  $\bar{\theta}_N$  we use Algorithm 1. The algorithm parameters are chosen following the recommendations provided in Section 3.3.1; we do not consider any warm-up iterations, and set  $\theta_0 = 0.01$ ,  $X_0^0 = y$ , for any  $n \in \mathbb{N}^*$ ,  $m_n = 1$ ,  $\delta_n = 10 \times n^{-0.8}/d$ ,  $\lambda = \min(5L^{-1}, \lambda_{\max})$  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 set  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 20$  burn-in iterations and compute  $\bar{\theta}_N$  using (3.5). To compute the MAP estimate we use SALSA with the following parameters: `inneriters` = 1, `outeriters` = 1000, `tol` =  $10^{-5}$  and `mu` =  $\bar{\theta}_N$ .

In Figure 4.7 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 4.8 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



**Figure 4.7** – Wavelet deconvolution with synthesis- $\ell_1$  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.

running average value  $(\bar{\theta}_N)_{N \in \mathbb{N}}$  throughout iterations.

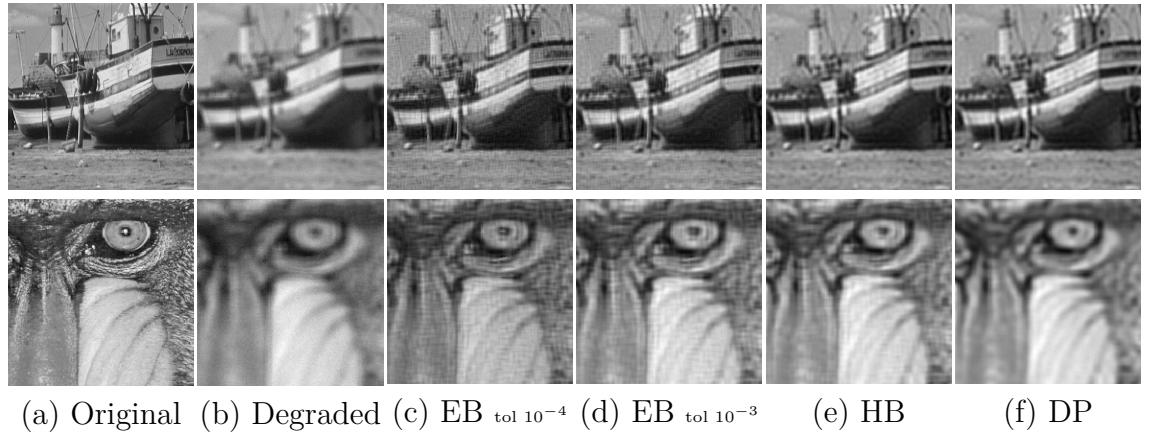


**Figure 4.8** – Wavelet deconvolution with synthesis- $\ell_1$  prior for `boat` image (SNR=20 dB). Evolution of (a) the iterates  $(\theta_n)_{n \in \mathbb{N}}$  in log-scale and (b) the relative change in  $(\bar{\theta}_N)_{N \in \mathbb{N}}$  for the proposed method.

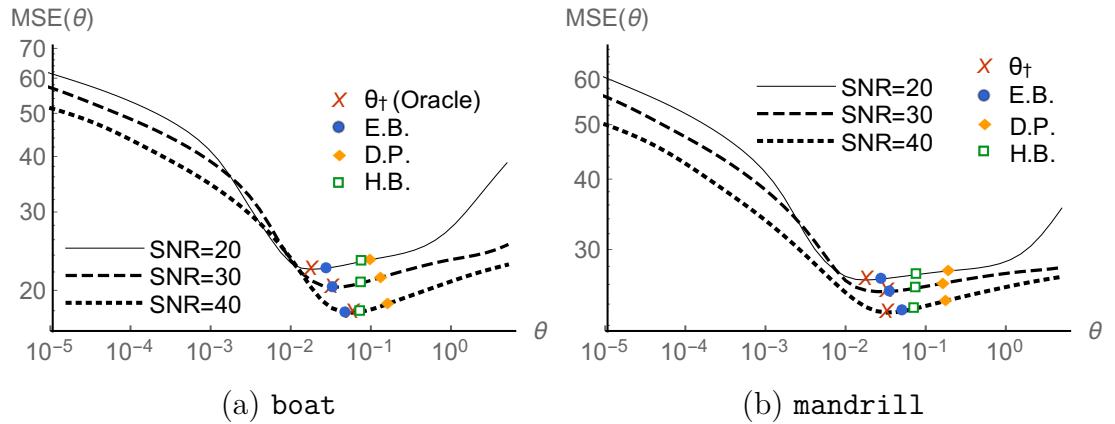
In Figure 4.9 we compare the results obtained by using each of the considered methods, showing a close-up on an image region that contains fine details and sharp edges. Figure 4.10 shows a plot of the MSE obtained with each method for the same two test images.

Table 4.3 shows the average MSE values and average computing times obtained for each method. 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.

For high SNR values, both Bayesian methods attain similar values of MSE, but



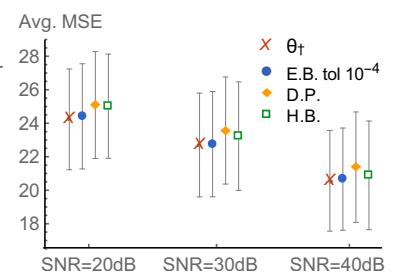
**Figure 4.9** – Wavelet deconvolution with synthesis- $\ell_1$  prior. Close-up on **boat** and **mandrill** images: (a) True image, (b) blurred and noisy (SNR=20 dB) observation  $y$ , (c)-(f) MAP estimators obtained with Empirical Bayes (tol.  $10^{-4}$  and  $10^{-3}$ ), hierarchical Bayes and discrepancy principle, respectively.



**Figure 4.10** – Wavelet deconvolution with synthesis- $\ell_1$  prior - Mean squared error (MSE) obtained for (a) **boat** and (b) **mandrill** for different values of  $\theta$ . We compare the values obtained with empirical Bayes with tolerance  $10^{-4}$ , discrepancy principle, hierarchical Bayes, and the optimal value  $\theta_{\dagger}$ .

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	MSE	Time	MSE	Time	MSE	Time
$\theta_{\dagger}$	24.23		22.70		20.56	
$tol 10^{-4}$ EB	<b>24.40</b>	4.48	<b>22.80</b>	3.59	<b>20.70</b>	2.44
$tol 10^{-3}$ EB	24.70	0.36	22.90	0.25	20.80	0.09
DP	25.09	13.93	23.57	28.64	21.38	61.03
HB	25.01	11.61	23.23	23.87	20.89	50.86

(a)



(b)

**Table 4.3** – Wavelet deconvolution with synthesis- $\ell_1$  prior. (a) Table with average mean squared error (MSE) obtained for ten images with different algorithms (results with lowest MSE in bold). Average execution times expressed in minutes. In (b) we summarise the content of the table and show the standard deviation with error bars.

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 [2]). We included a detailed analysis of this in Appendix B.

## 4.2 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 [75]. 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 [126]. In particular, linear unmixing techniques assume the availability of a library of spectral signatures and use the following model:

$$y = Ax + w \quad (4.5)$$

where  $y \in \mathbb{R}^{d_f \times d_p}$  is the hyperspectral image with  $d_f$  frequency channels and  $d_p$  pixels;  $x \in \mathbb{R}^{d_m \times d_p}$  is the fractional abundance matrix;  $A \in \mathbb{R}^{d_f \times d_m}$  is a dictionary of pure spectral signatures for  $d_m$  different materials; and  $w$  is a  $d_f \times d_p$  Gaussian random variables with zero mean and covariance matrix  $\sigma I_{d_y}$  and  $\sigma > 0$ . In [75], the unmixing problem is solved by using the regulariser  $g$  given for any  $x \in \mathbb{R}^{d_m} \times \mathbb{R}^{d_p}$  by

$$g(x) = (\text{TV}(x), \|x\|_1) \quad \text{s.t.} \quad x \geq 0, \quad (4.6)$$

which is associated with a two-dimensional regularisation parameter  $\theta = (\theta^{\text{TV}}, \theta^1) \in \mathbb{R}^2$ .  $\theta^{\text{TV}} \in \mathbb{R}$  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  $x \in \mathbb{R}^{d_m} \times \mathbb{R}^{d_p}$  by

$$\text{TV}(x) = \sum_{i=1}^{d_p} \sum_{j \in \mathcal{V}_i} \|x_i - x_j\|_1, \quad (4.7)$$

where for any  $i \in \{1, \dots, d_p\}$ ,  $x_i \in \mathbb{R}^{d_m}$  denotes the  $i$ -th image pixel and  $\mathcal{V}_i$  its vertical and horizontal neighbour pixels (the first-order neighbourhood).

Although this regulariser is not separable and we would therefore have to use Algorithm 3 with two MCMC chains, our empirical results suggest that it is possible to use a pseudo-likelihood approximation estimate  $\theta$  using a single MCMC chain together with the expression of  $\nabla_\theta \log Z(\theta)$  for the homogeneous case. The reason for doing this is twofold. First, with this approximation we can compare our results with the hierarchical Bayesian method from [105], which we would otherwise not be able to apply to this problem. Second, using Algorithm 3 with this particular application is very difficult due to numerical instabilities stemming from the way in which we implement the proximal operator with the SUNSAL solver [75]. Sampling from the prior becomes very difficult as the SUNSAL solver has not been designed to receive negative values in its input (which is often the situation given that we add Gaussian noise to the samples at every step) and this leads to numerical instabilities in the Markov chain sampling from the prior. For these reasons we have chosen to work under this approximation and use Algorithm 2.

More precisely, we consider  $[\partial \log Z / \partial \theta^1](\theta) = d/\theta^1$  and  $[\partial \log Z / \partial \theta^{TV}](\theta) = d/\theta^{TV}$ . Although  $x \mapsto \text{TV}(x)$  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. Notice that the dimension of each pseudo-separable component is set to  $d$  (i.e.  $d_1 = d$  and  $d_{\text{TV}} = d$ ). In standard cases, when using Algorithm 2, each separable component has a dimension smaller than  $d$  and acts on different components of  $x$  (this is explained in Section 3.1.2). In this experiment the regulariser is not truly separable due to the positivity constraint, so each regulariser term depends on the full vector  $x$ .

We consider the experiment A-*Simulated Data Sets* case 1) *Simulated Data Cube 1* presented in [75, 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  $d_p = 75 \times 75 = 5625$  pixels and  $d_f = 224$  frequency bands per pixel. For full details see [75]. We follow the exact same procedure as presented there, except for a modification in the spectral signature dictionary  $A$ . In [75] they consider a dictionary  $A \in \mathbb{R}^{224 \times 240}$ , which

is a library generated from a random selection of 240 materials from the USGS library<sup>2</sup>. Here we consider a simplified version where we only select  $d_m = 12$  random materials, thus having  $A \in \mathbb{R}^{224 \times 12}$ . 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 4.11 (only the 5 present end-members are shown)

We use the proposed algorithm to estimate  $\theta^{TV}$  and  $\theta^1$  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 [105] and by using the oracle value  $\theta_\dagger$  that maximises the estimation signal-to-reconstruction-error (SRE) given by  $\|x^0\|_2^2 / \|x^0 - \hat{x}_{MAP}\|_2^2$ .

We evaluated the proximal operator of  $x \mapsto \theta^{TV} TV(x) + \theta^1 \|x\|_1$  using SUNSAL solver from [75] with 20 iterations. We address the positivity constraint separately by using its Moreau-Yosida envelope (3.15), leading to the additional term  $x \mapsto (x - \Pi_+(x))/\lambda$  where  $\Pi_+$  is the projection operator onto  $[0, +\infty)^{d_m} \times [0, +\infty)^{d_p}$ , and  $\lambda$  is the same smoothing parameter used for the other proximal operators.

To speed up the convergence, we use a gradient preconditioning technique explained in Section 3.3.7. Since we use the preconditioned gradient of  $f_y$  instead of the gradient of  $f_y$ , the Lipschitz constant becomes  $L = 1/\sigma^2$ . The algorithm parameters are chosen following the recommendations provided in Section 3.3.1; we set  $\theta_0^1 = 10$ ,  $\theta_0^{TV} = 10$ , we initialised  $X_0^0$  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  $n \in \mathbb{N}^*$ ,  $m_n = 1$ ,  $\delta_n = n^{-0.8}/(d_p d_m)$ .

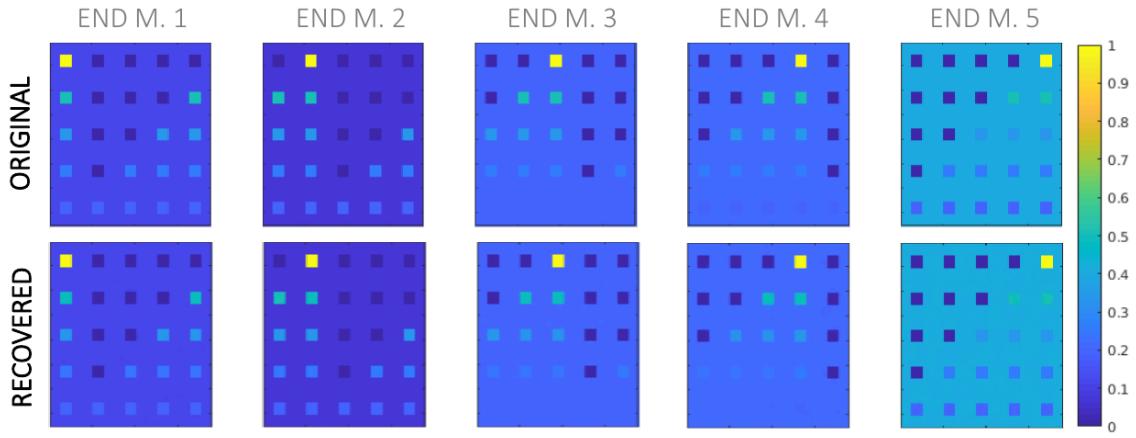
Special care was taken when setting  $\gamma > 0$  and  $\lambda > 0$  due to the preconditioning. We set  $\gamma = 1/(L + 2/\lambda)$  for any  $n \in \mathbb{N}$  and  $\lambda = 0.9 \times \lambda_A/L$ , where  $\lambda_A$  is the largest eigenvalue of  $(A^T A)^{-1}$ . We run the algorithm for 50 iterations and compute  $(\bar{\theta})_{N \in \mathbb{N}}$  as defined in (3.5) with  $(\omega_n)_{n \in \mathbb{N}}$  set to have  $N_0 = 30$  burn-in iterations.

In Figure 4.11 we display the MAP recovery of the synthetic fractional abundances using the estimated values of  $\theta^{TV}$  and  $\theta^1$  with the SUNSAL solver for SNR=30 dB.

Figure 4.12 (a) shows the evolution of the iterates  $(\theta_n^1)_{n \in \mathbb{N}}$  and  $(\theta_n^{TV})_{n \in \mathbb{N}}$  and Figure 4.12 (c) shows the relative change in the running averages  $(|\bar{\theta}_{N+1} - \bar{\theta}_N|/\bar{\theta}_N)_{N \in \mathbb{N}}$

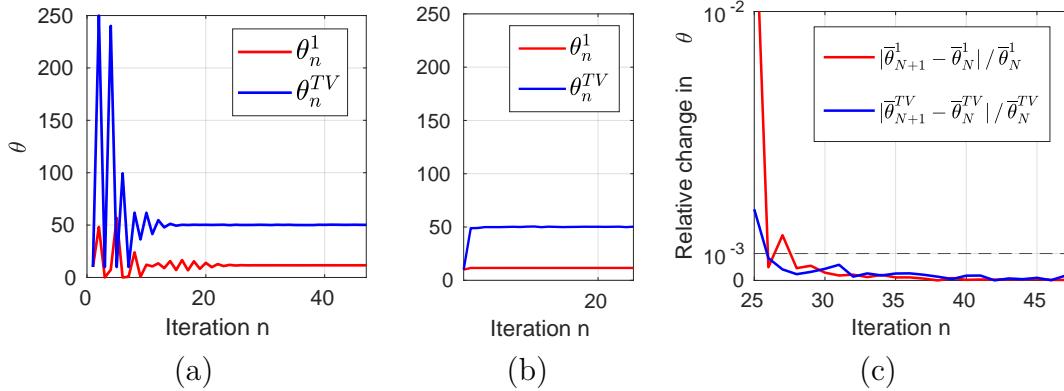
---

<sup>2</sup>Available online: <http://speclab.cr.usgs.gov/spectral.lib06>



**Figure 4.11** – Hyperspectral Unmixing - Synthetic fractional abundances for 5 end-members. Original and MAP estimates for SNR=30 dB using the empirical Bayes posterior (2.10).

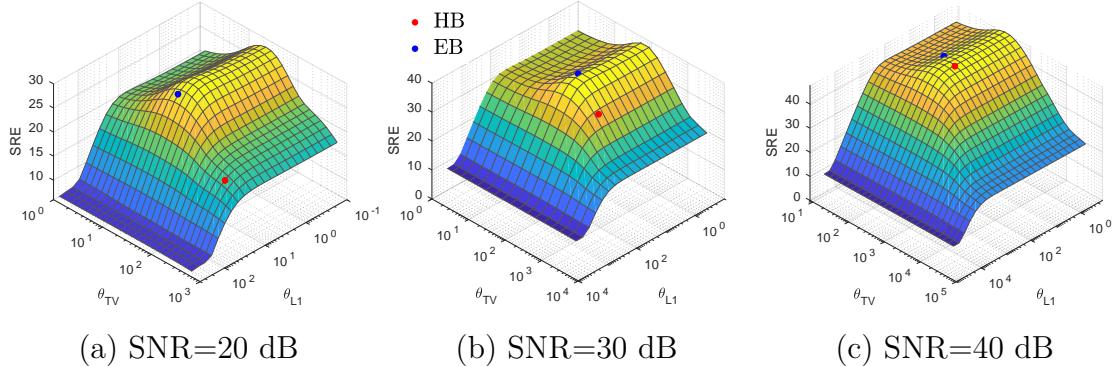
throughout iterations for SNR=30 dB. Observe the excellent convergence properties of the proposed scheme, which stabilises in as little as 25 iterations. Moreover, Figure 4.12 (b) shows the evolution of the iterates  $(\theta_n^1)_{n \in \mathbb{N}}$  and  $(\theta_n^{TV})_{n \in \mathbb{N}}$  obtained using Algorithm 5 instead of Algorithm 2 for the same experiment. We can see that the convergence is even faster for Algorithm 5 (where the exact maximisation step is used) and that both algorithms converge to the same estimates (see discussion in Section 3.1.5).



**Figure 4.12** – Hyperspectral Unmixing with SNR=30 dB - Evolution of the iterates  $(\theta_n^1)_{n \in \mathbb{N}}$  and  $(\theta_n^{TV})_{n \in \mathbb{N}}$  using (a) Algorithm 2 and (b) Algorithm 5. In (c) we show the relative successive differences  $(|\bar{\theta}_{N+1} - \bar{\theta}_N| / \bar{\theta}_N)_{N \in \mathbb{N}}$  for Algorithm 2, where the relative change is computed after 25 burn-in iterations.

The obtained results are reported in Table 4.4 and summarised in Figure 4.13, 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 2.3.3, 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  $\theta$  assigns a high weight to values of  $\theta$  that lead to



**Figure 4.13** – 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 [105].

Method	Stop criteria	SNR=20 dB		SNR=30 dB		SNR=40 dB	
		SRE	Time (s)	SRE	Time (s)	SRE	Time (s)
$\theta_{\dagger}$ (Oracle)	–	29.38	–	38.61	–	47.64	–
EB	50 iters.	27.46	36	38.42	37	45.68	42
HB [105]	15 iters.	18.33	76	31.72	77	47.36	76

**Table 4.4** – Hyperspectral unmixing - Signal to reconstruction error (SRE) obtained for different SNR values along with computing times expressed in seconds.

bad models, *i.e.* a misspecified prior  $x \mapsto p(x|\theta)$ , the impact of this misspecification on the recovered estimates depends on the degree of concentration of 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 4.4 that the computing times for the empirical Bayesian method are approximately two times faster than the ones for the hierarchical method.

## 4.3 Denoising with a total generalised 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  $y \sim \mathcal{N}(x, \sigma^2 I_{d_y})$  with  $\sigma^2 > 0$  and where we use the following prior

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

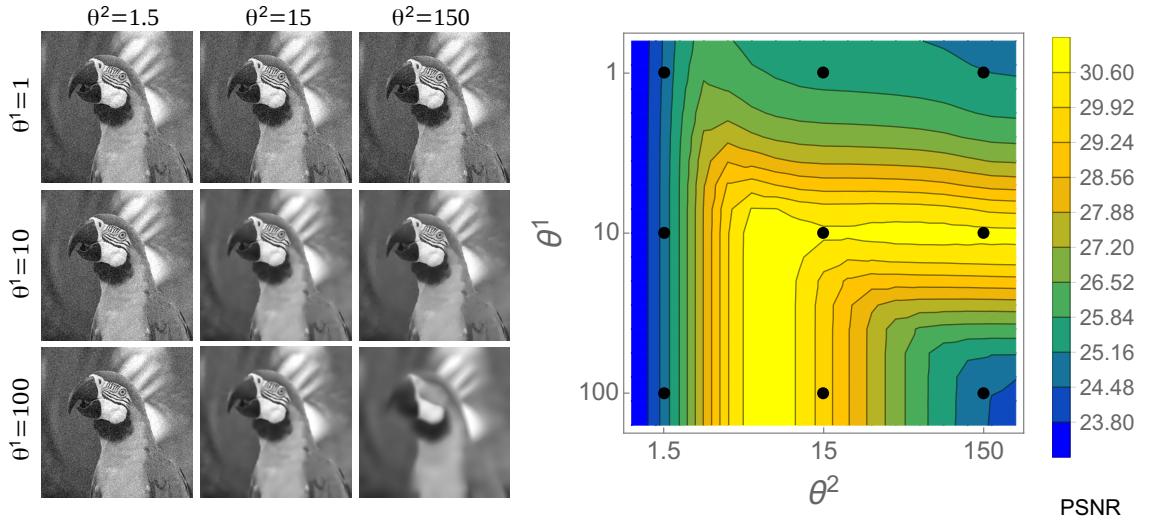
where  $\varepsilon > 0$  and where  $\text{TGV}_{\theta^1, \theta^2}^2(x)$  is a second-order generalisation of the conventional total variation regulariser, given, for any  $(\theta^1, \theta^2) \in [0, +\infty)^2$  and  $x \in \mathbb{R}^d$ , by

$$\text{TGV}_{\theta^1, \theta^2}^2(x) = \min_{r \in \mathbb{R}^{2d}} \{\theta^1 \|r\|_{1,2} + \theta^2 \|J(\Delta x - r)\|_{1, \text{Frob.}}\}. \quad (4.8)$$

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) [43]. This generalisation was first considered in [34] and further studied in [18] 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  $\theta^1$  and  $\theta^2$ , which control the strength as well as the characteristics of the regularisation enforced (as explained in [43], the TGV regularisation behaves like the standard TV regularisation for large  $\theta^2$  values, whereas for small values it behaves like the  $\ell_1$ -Frobenius norm of the discrete image-Hessian). Figure 4.14 below illustrates the dramatic effect that these two parameters have on the quality of the recovered MAP estimate. Observe the strong coupling between  $\theta^1$  and  $\theta^2$ , which makes setting their values particularly challenging.

However, this prior is not in the exponential family because  $\theta^1$  and  $\theta^2$  play a role in the definition of the statistic  $\text{TGV}_{\theta^1, \theta^2}^2(x)$ . Therefore, our methodology and theory do not directly apply. Also note that the additional regularisation  $\varepsilon \|x\|_2^2$  with  $\varepsilon > 0$



**Figure 4.14** – Denoising with TGV prior. MAP estimates for different values of  $\theta^1$  and  $\theta^2$  for parrot image with SNR = 5.6 dB (left). PSNR for different values of  $\theta^1$  and  $\theta^2$  (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.

is necessary to guarantee that the prior distribution is proper<sup>3</sup>, 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 (4.8) is not separable and homogeneous). We use  $\varepsilon = 10^{-10}$ .

In order to apply the proposed methodology to the estimation of  $\theta^1$  and  $\theta^2$  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  $\theta^1, \theta^2 > 0$

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

with

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

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

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

---

<sup>3</sup>This additional regularisation was not necessary in the total variation experiment in Section 4.1.1 as Algorithm 1 does not require sampling from the prior distribution, so improper priors can be used.

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

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

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

This approximation of the gradient arises from omitting the terms

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

and

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

in the calculation of the partial derivatives  $\frac{\partial}{\partial \theta^1} \log p(x|\theta^1, \theta^2)$  and  $\frac{\partial}{\partial \theta^2} \log p(x|\theta^1, \theta^2)$ .

The omission of these terms is necessary because they are not directly available and would require being separately estimated at every iteration (in contrast, the terms that we do include in the gradient approximation,  $g_1$  and  $g_2$ , can be obtained as a by-product of the proximal operator evaluation). Although this approximation introduces an additional bias in the stochastic gradients driving Algorithm 3<sup>4</sup>, 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  $\theta^1$  and  $\theta^2$ .

In our experiments, we implement Algorithm 3 with this approximate gradient and follow the recommendations provided in Section 3.3.1 to set the algorithm parameters; we perform 25 warm-up iterations and set  $\theta_0^1 = \theta_0^2 = 10$ ,  $X_0^0 = \bar{X}_0^0 = y$ , for any  $n \in \mathbb{N}^*$ ,  $m_n = 1$ ,  $\delta_n = 20 \times n^{-0.8}/d$ , and we set  $\lambda = \min(5L^{-1}, \lambda_{\max})$  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  $\|\bar{\theta}_{N+1} - \bar{\theta}_N\|_\infty \leq 10^{-4}$  and iii)  $\|\bar{\theta}_{N+1} - \bar{\theta}_N\|_\infty \leq 10^{-3}$ . Again, we compute  $\bar{\theta}_N$  using (3.5), setting  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 20$  burn-in iterations.

We also considered a thinning of 6 iterations in the chain associated with the

---

<sup>4</sup>A rigorous analysis of this bias should also consider the points where  $\text{TGV}_{\theta^1, \theta^2}^2(x)$  is not differentiable w.r.t.  $\theta^1$  and  $\theta^2$ . This can be achieved by using similar techniques to [46].

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 Section 3.3.5). To compute the  $\text{TGV}_{\theta^1, \theta^2}^2$  norm and proximal operator, we use the iterative primal-dual algorithm [43].

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  $255 \times 255$  pixels, and then use the estimated  $\theta^1$  and  $\theta^2$  values to compute the MAP estimate of the entire image<sup>5</sup>. We consider the same ten test images used in Section 4.1 and we set the noise variance  $\sigma^2$ , 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  $\theta^1$  and  $\theta^2$  and then use them to compute the MAP estimator  $\hat{x}_{\text{MAP}}$  (given by (1.7)) using the same solver [43] we use for the proximal operator. We measure estimation performance by computing the peak-signal-to-noise-ratio (PSNR) given by  $\text{PSNR}(x, \hat{x}_{\text{MAP}}) = -10 \log_{10} \|x - \hat{x}_{\text{MAP}}\|_2^2/d$ . All the PSNR plots shown in Figure 4.17, Figure 4.18 and Figure 4.21 were computed with the entire images instead of the cropped patches.

Table 4.5 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 4.15 depicts the original image, the noisy observation and the recovered MAP estimates for the `boat` and `lake` test images for the case with  $\text{SNR} = 8$  dB. In the `boat` image we can see some denoising artefacts in the sky. This is mostly due to the choice of prior distribution and not to incorrectly tuned regularisation parameters (see Figure 4.17). In fact, the reconstructions

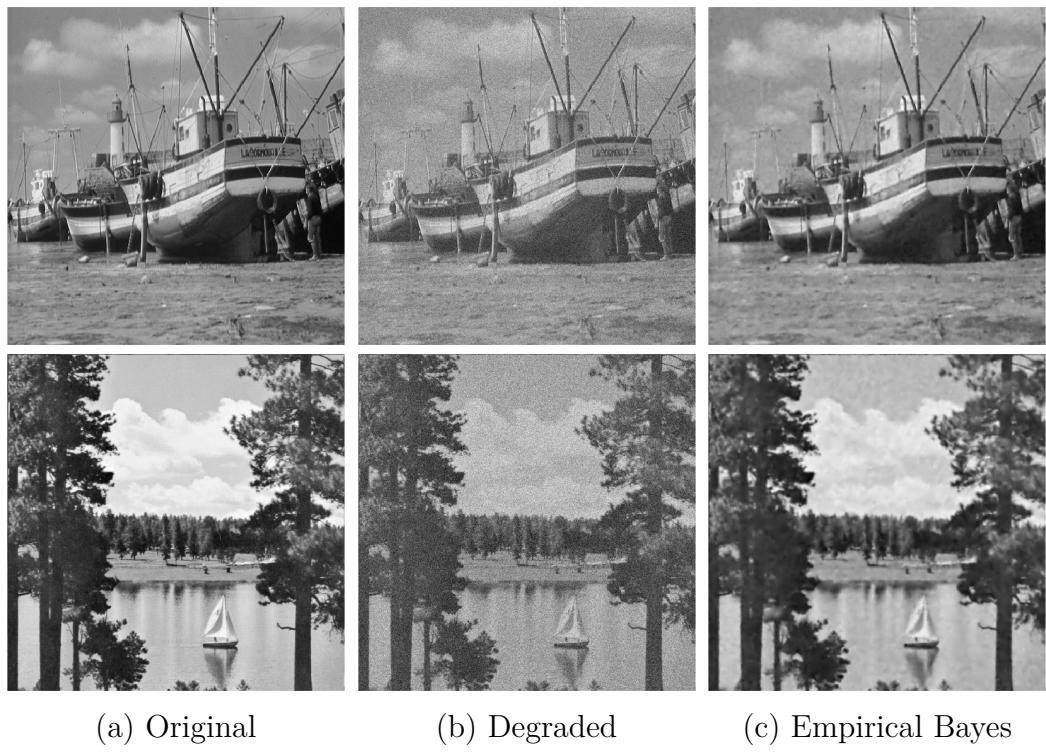
---

<sup>5</sup>For homogeneous regularisers,  $\theta$  is asymptotically independent of the dimension of  $x$  when  $d$  is large [105], 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.

Method	SNR=8 dB		SNR=12 dB		SNR=20 dB	
	PSNR	Time	PSNR	Time	PSNR	Time (min)
$\theta_\dagger$ (Oracle)	$27.80 \pm 2.35$		$30.21 \pm 2.12$		$35.60 \pm 1.77$	
2000 iter EB	<b><math>27.11 \pm 2.81</math></b>	131.10	$29.69 \pm 2.33$	96.41	$35.48 \pm 1.81$	95.06
$tol 10^{-4}$ EB	$27.09 \pm 2.84$	24.61	<b><math>29.72 \pm 2.33</math></b>	23.27	$35.47 \pm 1.81$	44.70
$tol 10^{-3}$ EB	$27.00 \pm 2.96$	3.04	$29.50 \pm 2.71$	2.18	<b><math>35.57 \pm 1.79</math></b>	5.03

**Table 4.5** – Denoising with TGV prior. Average mean squared error  $\pm$  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  $\theta^1$  and  $\theta^2$  estimates.

obtained with the optimal value  $\theta_\dagger$  that maximises the PSNR preserve more fine details but display even more pronounced artefacts as shown in Figure 4.16. For better reconstructions a different prior distribution should be used.



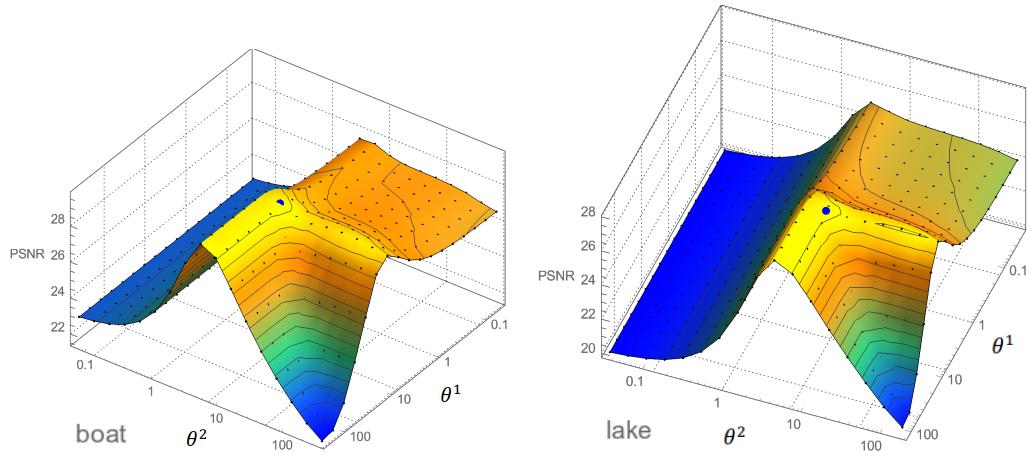
**Figure 4.15** – Denoising with TGV prior for `boat` and `lake` test images: (a) True image, (b) noisy observation  $y$  (SNR=8 dB), (c) MAP estimators obtained with EB. We show the full image (not the patches).

More interestingly, Figure 4.17 shows the landscape of the PSNR as a function of  $\theta^1$  and  $\theta^2$  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



**Figure 4.16** – Denoising with TGV prior for `boat` and `lake` test images: MAP estimators obtained with the optimal value  $\theta_{\dagger}$  that maximises the PSNR.

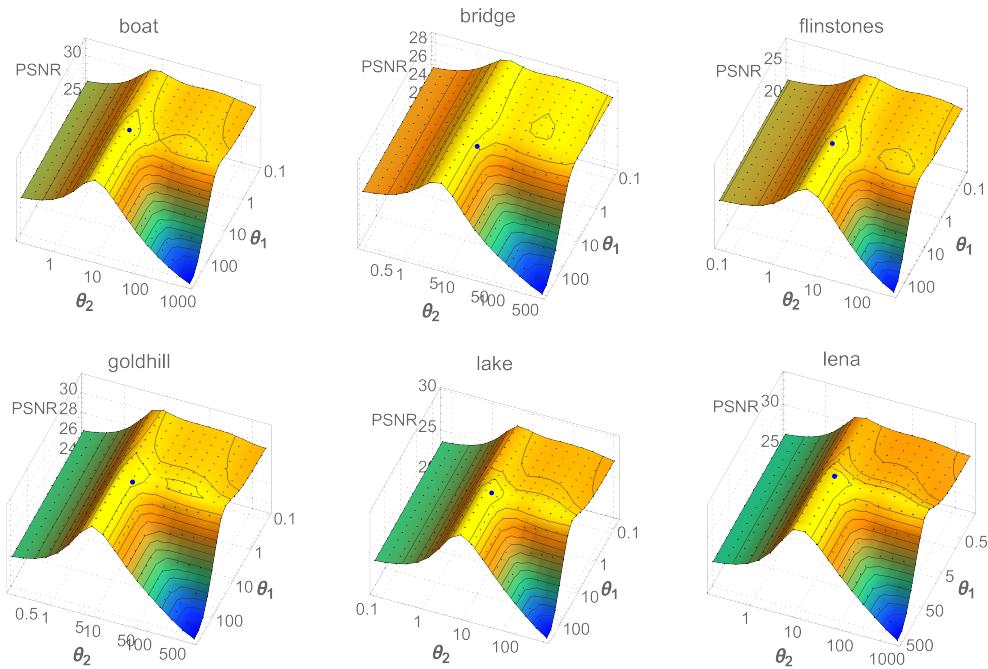
truth.



**Figure 4.17** – Denoising with TGV prior on `boat` and `lake` images (SNR=8 dB). PSNR for different values of  $\theta^1$  and  $\theta^2$ . Blue marker shows the location of  $\bar{\theta}_N$  estimated with empirical Bayes using 2000 iterations. Associated images shown in Figure 4.15.

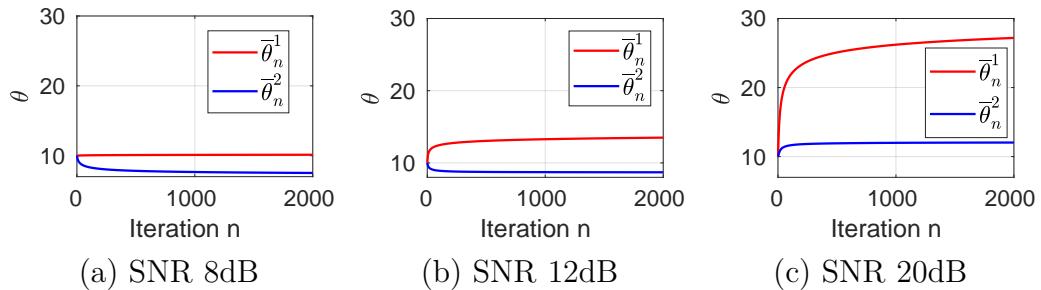
The PSNR surfaces in Figure 4.17, Figure 4.18 and Figure 4.21, were computed by evaluating the solver in multiple points (shown as a grid of grey dots on each plot) and then interpolating to show the full surface. For this reason the level-set curves (plotted as grey curves on the surface) are simple numerical approximations to the real level-sets and the sharp peaks or abrupt changes observed in them are just a product of the numerical approximation. In Figure 4.17 both images seem to present two modes in the PSNR surface. This multi-modality was observed in most of the considered images for all SNR levels, see for example Figure 4.18.

Following on from this, Figure 4.19 and Figure 4.20 show respectively the evolution of the iterates and the relative change in the estimated values of  $\theta^1$  and  $\theta^2$ , for



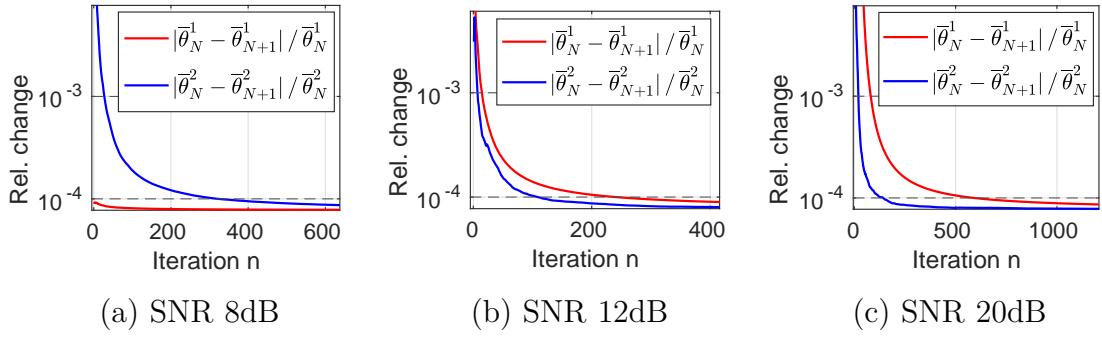
**Figure 4.18** – Denoising with TGV prior for 6 test images (SNR=12 dB). PSNR for different values of  $\theta^1$  and  $\theta^2$ . Blue marker shows the location of  $\bar{\theta}_N$  estimated with empirical Bayes using  $tol = 10^{-4}$  as a stopping criterion.

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.



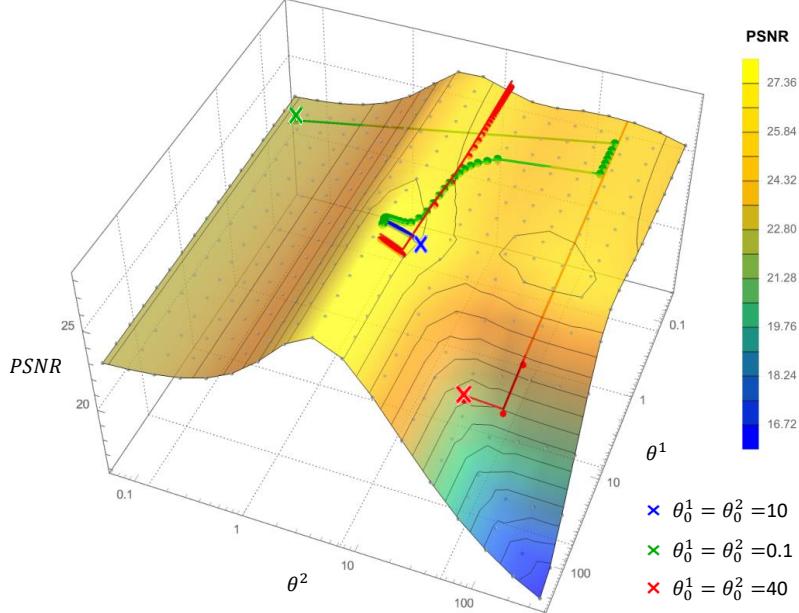
**Figure 4.19** – Denoising with TGV prior. Evolution of the iterates  $(\theta_n^1)_{n \in \mathbb{N}}$  and  $(\theta_n^2)_{n \in \mathbb{N}}$  for the `lake` test image for different SNR values.

Lastly, Figure 4.21 below explores the robustness to different initialisations by showing the evolution of the iterates on the landscape of PSNR values for the `flintstones` 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  $\theta^1$  and  $\theta^2$  that are close-to-optimal in terms of the resulting PSNR. In Figure 4.22 we show the MAP estimates



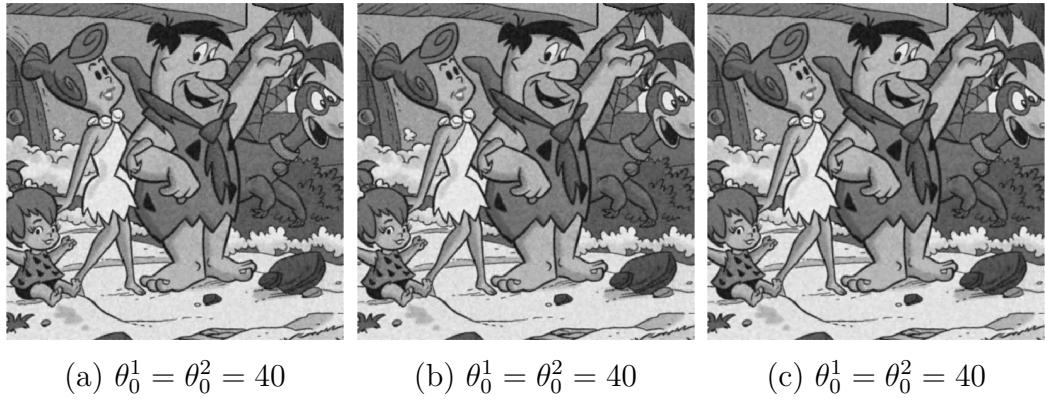
**Figure 4.20** – Denoising with TGV prior. Relative successive differences  $|\bar{\theta}_N^i - \bar{\theta}_{N+1}^i|/\bar{\theta}_N^i$  with  $i = 1, 2$  for the proposed method with the `lake` test image for different SNR values.

computed with each one of the three estimated  $\bar{\theta}_N$  from Figure 4.21. As it may be seen, the perceptual difference between the MAP estimates is negligible.



**Figure 4.21** – Denoising with TGV prior on the `flintstones` image (SNR=12 dB). Evolution of the iterates  $(\theta_n^1)_{n \in \mathbb{N}}$  and  $(\theta_n^2)_{n \in \mathbb{N}}$  for different initial values  $\theta_0^1$  and  $\theta_0^2$ . When initialising with  $\theta_0^1 = \theta_0^2 = 40$  (red) the algorithm converges to a different point with a similar PSNR.

Nevertheless, 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  $\theta^1$  with  $\theta^2$  fixed to a small value - e.g.  $\theta^2 = 1$  - to keep the model



**Figure 4.22** – Denoising with TGV prior for `flintstones` test image with SNR=12 dB: MAP estimators obtained with three different estimations of  $\bar{\theta}_N$  obtained by running the EB algorithm with three different initial values  $\theta_0^1$  and  $\theta_0^2$ .

close to the conventional total variation regulariser, and then update both  $\theta^1$  with  $\theta^2$  until the convergence criterion is satisfied.

To conclude, we note that there are several other generalisations of the total variation regularisation (see [18]). We have chosen to perform our experiments with (4.8) because of the availability of the efficient MATLAB implementation [43]. 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 [18] that is very similar to (4.8).

# Chapter 5

## Beyond imaging applications

In the previous chapter, we focused on the problem of estimating regularisation parameters in imaging problems. In this chapter we want to show the scope of the proposed methodology for estimating other kinds of parameters in other types of inverse problems.

First, in Section 5.1, we re-introduce the algorithm in more general terms so that it can be applied to a broader range of maximum likelihood estimation problems involving intractable likelihood functions.

We then demonstrate the generalised methodology with three experiments that involve a variety of unknown model parameters. Section 5.2 presents an application to empirical Bayesian logistic regression, where the goal is to estimate a hyperparameter from the prior distribution of the regression coefficients. In Section 5.3 we consider a challenging application related to audio compressed sensing analysis, where we use the proposed methodology to estimate a regularisation parameter that controls the degree of sparsity enforced. Finally, Section 5.4 presents an application to a high-dimensional empirical Bayesian logistic regression with random effects for which the optimisation problem (5.2) is not convex. All experiments were carried out on an Intel i9-8950HK@2.90GHz workstation running MATLAB R2018a.

### 5.1 Generalised SAPG algorithm

In Chapter 3 we considered imaging models of the form  $p(x, y|\theta) \propto \exp[-f_y(x) - \theta^T g(x)]$  and used Fisher's identity to write the intractable gradient of the marginal likeli-

hood  $\log p(y|\theta)$  in terms of an expectation  $\nabla_\theta \log p(y|\theta) = \mathbb{E}_{x|y,\theta}\{\nabla_\theta \log p(x,y|\theta)\}$ . This enabled us to obtain a Monte Carlo estimate of  $\nabla_\theta \log p(y|\theta)$  by computing

$$\Delta_{m,\theta} = \frac{1}{m} \sum_{k=1}^m \nabla_\theta \log p(X_k, y|\theta) = -\frac{1}{m} \sum_{k=1}^m g(X_k) - \nabla_\theta \log Z(\theta), \quad (5.1)$$

where  $(X_k)_{k \in \{0, \dots, m\}}$  was 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.

In this chapter we focus on more general models of the form  $p(x, y|\theta) \propto \exp[-f(x, y, \theta)]$  where  $f$  is convex w.r.t.  $x$  and the gradient  $\nabla_\theta \log p(x, y|\theta)$  is not necessarily given by  $-g(x) - \nabla_\theta \log Z(\theta)$ . Moreover, we also consider a more general version of Equation (2.8) where the maximum likelihood estimator is given by

$$\theta^* \in \arg \max_{\theta \in \Theta} \log p(y|\theta) - \varphi(\theta), \quad (5.2)$$

allowing the use of a penalty function  $\varphi : \Theta \rightarrow \mathbb{R}$ , or set  $\varphi = 0$  to recover the standard maximum likelihood estimator. We can then consider the following recursion for any  $n \in \mathbb{N}$

$$\theta_{n+1} = \Pi_\Theta[\theta_n + \delta_{n+1} \{\Delta_{m_n, \theta_n} - \nabla_\theta \varphi(\theta_n)\}], \quad \Delta_{m_n, \theta_n} = \frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_k^n, y|\theta_n), \quad (5.3)$$

which leads to Algorithm 7 below.

---

**Algorithm 7** SAPG algorithm - General form
 

---

```

1: Input: initial  $\{\theta_0, X_0^0\}$ ,  $(\delta_n, \omega_n, m_n)_{n \in \mathbb{N}}$ ,  $\Theta$ , kernel parameters  $\gamma, \lambda$ , iterations  $N$ .
2: for  $n = 0$  to  $N - 1$  do
3:   if  $n > 0$  then
4:     Set  $X_0^n = X_{m_{n-1}}^{n-1}$ ,
5:   end if
6:   for  $k = 0$  to  $m_n - 1$  do
7:     Sample  $X_{k+1}^n \sim R_{\gamma, \lambda, \theta_n}(X_k^n, \cdot)$ ,
8:   end for
9:   Set  $\Delta_{m_n, \theta_n} = \frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_k^n, y|\theta_n)$ 
10:  Set  $\theta_{n+1} = \Pi_\Theta[\theta_n + \delta_{n+1} \{\Delta_{m_n, \theta_n} - \nabla_\theta \varphi(\theta_n)\}]$ .
11: end for
12: Output:  $\bar{\theta}_N$  computed with (3.5).
    
```

---

Having defined Algorithm 7 in this general form, we can now demonstrate the proposed methodology with a broader range of estimation problems. A detailed theoretical analysis of this generalised version of the algorithm for smooth cases (i.e. when  $f$  is continuously

differentiable) is available in [46].

We want to point out that some of the implementation guidelines provided in Section 3.3 might not directly apply to Algorithm 7. In particular, notice that the  $\Delta_{m,\theta}$  is no longer given by  $-\frac{1}{m} \sum_{k=1}^m g(X_k) - \nabla_\theta \log Z(\theta)$ , so it needs to be carefully derived for each problem. More importantly, in most experiment in Chapter 4 the Lipschitz constant of  $\nabla_x \log p(x|y, \theta)$  was upper bounded by  $1/(L_y + \lambda^{-1})$ , so we suggested setting  $\gamma = 0.98(L_y + 1/\lambda)^{-1}$ . This no longer applies to Algorithm 7 as  $\nabla_x \log p(x|y, \theta)$  can have different forms and is not necessarily upper bounded by  $1/(L_y + \lambda^{-1})$ . Therefore the step-size  $\gamma$  needs to be set to be smaller than the inverse of the Lipschitz constant of  $\nabla_x \log p(x|y, \theta)$ , and this Lipschitz constant needs to be calculated (or at least upper bounded) for each particular problem.

## 5.2 Bayesian Logistic Regression

In this first experiment we illustrate the proposed methodology with an empirical Bayesian logistic regression problem [112, 139]. We observe a set of covariates  $\{v_i\}_{i=1}^{d_y} \in \mathbb{R}^d$ , and binary responses  $\{y_i\}_{i=1}^{d_y} \in \{0, 1\}$ , which we assume to be conditionally independent realisations of a logistic regression model: for any  $i \in \{1, \dots, d_y\}$ ,  $y_i$  given  $\beta$  and  $v_i$  has distribution  $\text{Ber}(s(v_i^\top \beta))$ , where  $\beta \in \mathbb{R}^d$  is the regression coefficient,  $\text{Ber}(\alpha)$  denotes the Bernoulli distribution with parameter  $\alpha \in [0, 1]$  and  $s(u) = e^u / (1 + e^u)$  is the cumulative distribution function of the standard logistic distribution. The prior for  $\beta$  is set to be  $N(\theta \mathbf{1}_d, \sigma^2 \mathbf{I}_d)$ , the  $d$ -dimensional Gaussian distribution with mean  $\theta \mathbf{1}_d$  and covariance matrix  $\sigma^2 \mathbf{I}_d$ , where  $\theta$  is the parameter we seek to estimate,  $\mathbf{1}_d = (1, \dots, 1) \in \mathbb{R}^d$ ,  $\sigma^2 = 5$  and  $\mathbf{I}_d$  is the  $d$ -dimensional identity matrix<sup>1</sup>. Following an empirical Bayesian approach, the parameter  $\theta$  is computed by maximum marginal likelihood estimation using Algorithm 7 with the marginal likelihood given by

$$p(y|\theta) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \left\{ \prod_{i=1}^{d_y} s(v_i^\top \beta)^{y_i} (1 - s(v_i^\top \beta))^{1-y_i} \right\} e^{-\frac{\|\beta - \theta \mathbf{1}_d\|^2}{2\sigma^2}} d\beta. \quad (5.4)$$

Lemma 7 in Appendix A of [46] shows that (5.4) is log-concave with respect to  $\theta$ . We use the proposed methodology to estimate  $\theta_*$  for the Wisconsin Diagnostic Breast Cancer

---

<sup>1</sup>The exact value of  $\sigma$  is not crucial as long as it is within a range where it can be considered non-informative. As a rule of thumb, if changing the value of  $\sigma$  has a strong impact on the results, then it is probably not large enough. Common ranges go from 5 to 100.

dataset<sup>2</sup>, for which  $d_y = 683$  and  $d = 10$ , and where we normalise the covariates. In order to assess the quality of our estimation results, we also calculate  $p(y|\theta)$  over a grid of values for  $\theta$  by using a truncated harmonic mean estimator.

To implement Algorithm 7 we derive the log-likelihood function

$$\log p(y|\beta, \theta) = \sum_{i=1}^{d_y} \left\{ y_i v_i^T \beta - \log(1 + e^{(v_i^T \beta)}) \right\}, \quad (5.5)$$

and obtain the following expressions for the gradients used in the MCMC steps (3.14) and in the SA step on Line 10 in Algorithm 7, respectively

$$\nabla_\beta \log p(\beta|y, \theta) = \sum_{i=1}^{d_y} \left\{ y_i v_i - s(v_i^T \beta) v_i \right\} - \frac{(\beta - \theta \mathbf{1}_d)}{\sigma^2}, \quad (5.6)$$

$$\nabla_\theta \log p(\beta, y|\theta) = \langle \mathbf{1}_d, \beta - \theta \mathbf{1}_d \rangle / \sigma^2. \quad (5.7)$$

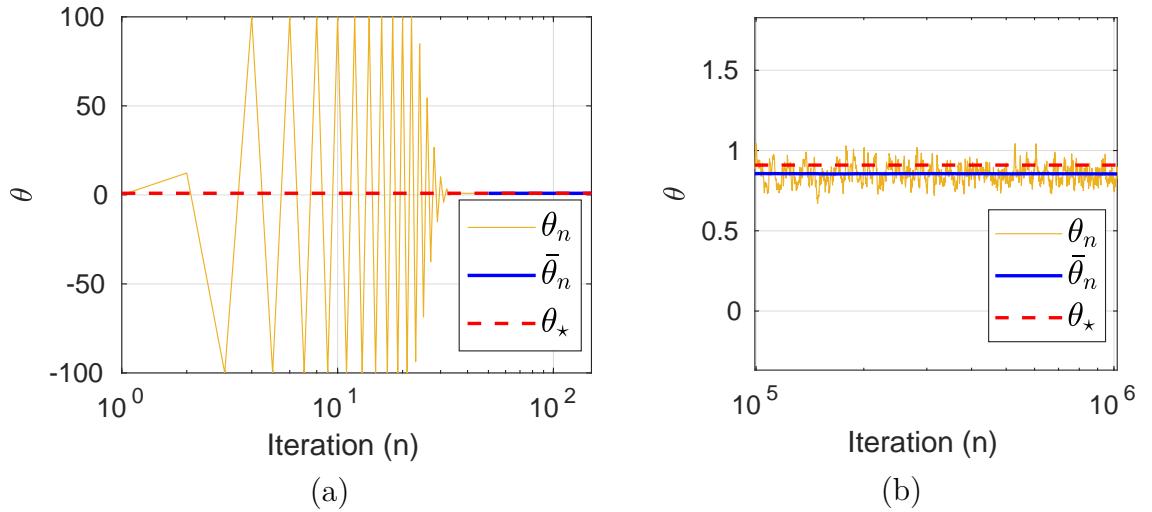
For the MCMC steps, we use a fixed step-size  $\gamma = 8.34 \times 10^{-5}$ , and batch size  $m_n = 1$ , for any  $n \in \mathbb{N}$ . On the other hand we consider, for the SA steps, the sequence of step-sizes  $\delta_n = 60/n^{-0.8}$ ,  $\varphi(\theta) = 0$ ,  $\Theta = [-100, 100]$  and  $\theta_0 = 0$ . Finally, we first run 100 iterations with fixed  $\theta_n = \theta_0$  to warm-up the Markov chain, and then run  $N = 10^6$  iterations of Algorithm 7, setting  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 50$  burn-in iterations, and compute  $\bar{\theta}_N$  using (3.5).

Figure 5.1 (a) shows the evolution of the iterates  $\theta_n$  during the first 100 iterations. Observe that the sequence initially oscillates, and then stabilises close to  $\theta_*$  after approximately 50 iterations. Figure 5.1 (b) presents the iterates  $\theta_n$  for  $n = 10^5, \dots, 10^6$ . For completeness, Figure 5.2 shows the histograms corresponding to the marginal posteriors  $p(\beta_j|y, v, \bar{\theta}_N)$ , for  $j = 1, \dots, 10$ , obtained as a by-product of Algorithm 7. In order to verify that the obtained estimate  $\bar{\theta}_N$  is close to the true MLE  $\theta_*$  we use a truncated harmonic mean estimator (THME) [120] to calculate the marginal likelihood  $p(y|\theta)$  for a range of values of  $\theta$ . Although obtaining the THME is usually computationally expensive, it is viable in this particular experiment as  $\beta$  is low-dimensional. More precisely, given  $n$  samples  $(\beta_i)_{i \in \{1, \dots, n\}}$  from  $p(\beta|y, \theta)$ , we obtain an approximation of  $p(y|\theta)$  by computing

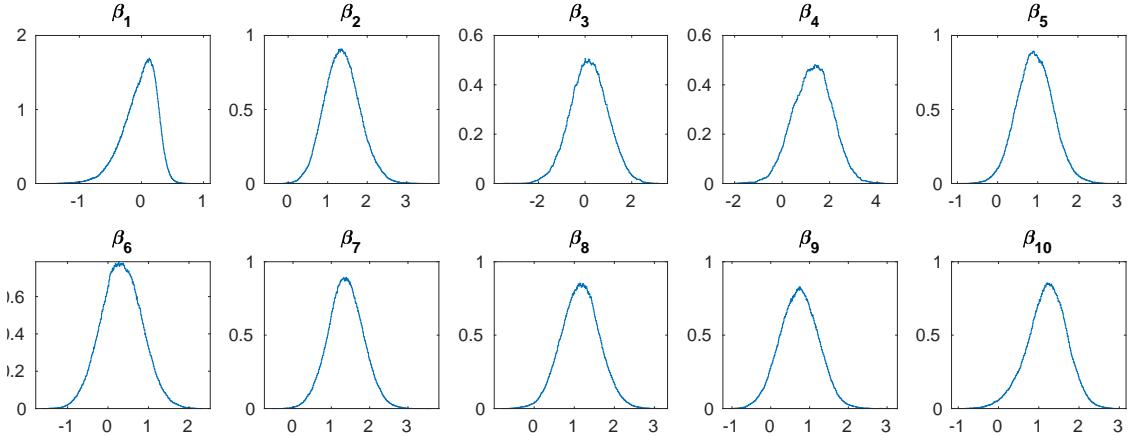
$$\hat{p}(y|\theta) = n \text{Vol}(\mathcal{A}) \left/ \left( \sum_{k=1}^n \frac{\mathbb{1}_{\mathcal{A}}(\beta_k)}{p(\beta_k, y|\theta)} \right) \right., \quad (5.8)$$

---

<sup>2</sup>Available online: [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))



**Figure 5.1** – Bayesian logistic regression - Evolution of the iterates  $\bar{\theta}_n$  and  $\theta_n$  for the proposed method during (a) burn-in phase and (b) convergence phase. An estimate of  $\theta_*$ , the true maximiser of  $p(y|\theta)$ , is plotted as a reference.

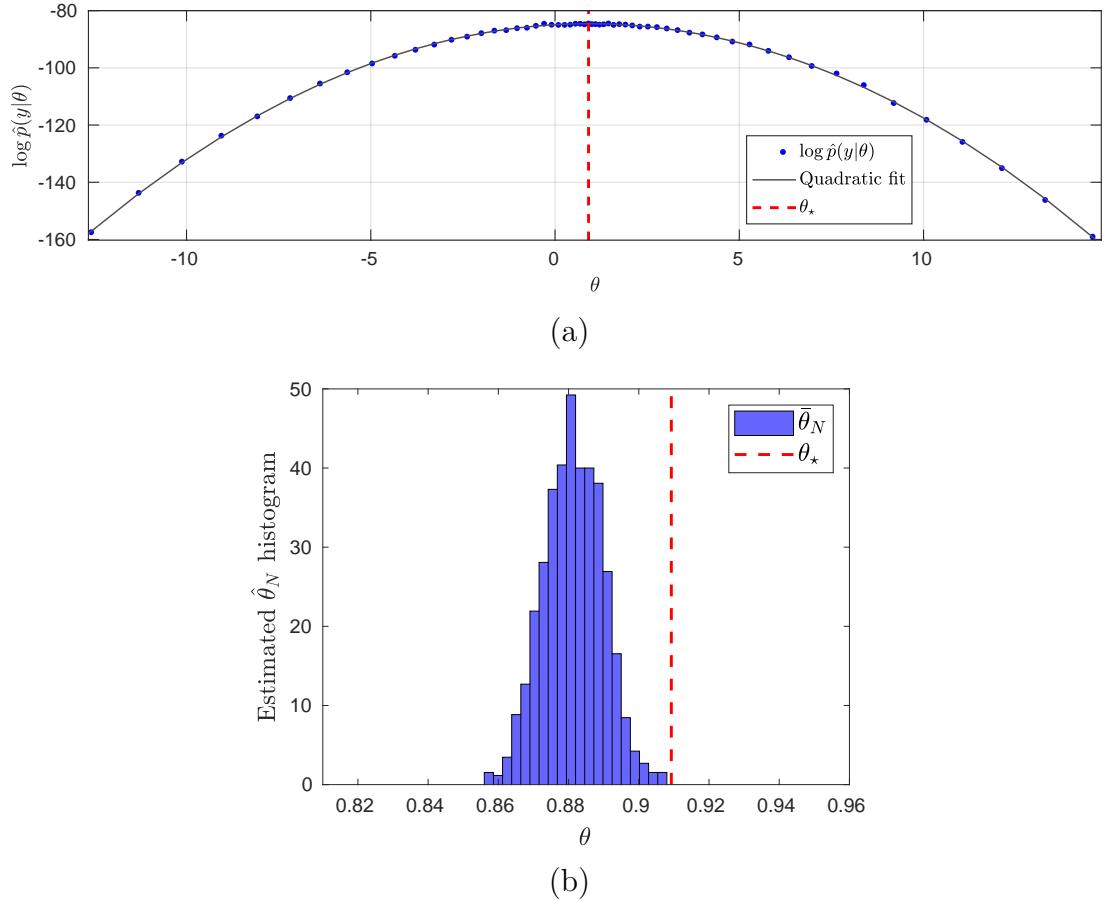


**Figure 5.2** – Bayesian logistic regression - Normalised histograms of each component of  $\beta$  obtained with  $2 \times 10^6$  Monte Carlo samples.

where  $\mathbf{A}$  is a  $d$ -dimensional ball centred at the posterior mean  $\bar{\beta} = n^{-1} \sum_{k=1}^n \beta_k$ , and with radius set such that  $n^{-1} \sum_{i=1}^n \mathbb{1}_{\mathbf{A}}(\beta_i) \approx 0.4$ . Using  $n = 6 \times 10^5$  samples, we obtain the approximation shown in Figure 5.3 (a), where in addition to the estimated points we also display a quadratic fit (corresponding to a Gaussian fit in linear scale), which we use to obtain an estimate of  $\theta_*$ .

To empirically study the estimation error involved, we replicate the experiment  $10^3$  times. Figure 5.3 (b) shows the obtained histogram of  $\{\bar{\theta}_{N,i}\}_{i=1}^{1000}$ , where we observe that all these estimators are very close to the true maximiser  $\theta_*$ . Note that the distribution of the estimation error is close to a Gaussian distribution, as expected for a maximum likelihood estimator. There is a small estimation bias of the order of 3%, which can be

attributed to the bias discussed in Section 3.1.4, and potentially to a small error in the estimation of  $\theta_*$ .



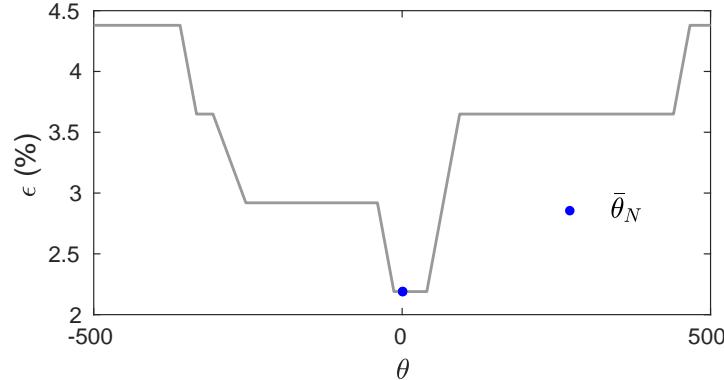
**Figure 5.3** – Bayesian logistic regression - (a) Estimated points of the marginal log-likelihood  $\log \hat{p}(y|\theta)$  with quadratic fit (corresponding to a Gaussian fit in linear scale). (b) Normalised histogram of  $\bar{\theta}_N$  for 1000 repetitions of the experiment. An estimate of  $\theta_*$ , the maximiser of  $\hat{p}(y|\theta)$ , is plotted as a reference.

We conclude this experiment by using Algorithm 7 to perform a predictive empirical Bayesian analysis on the binary responses. We split the original dataset into an 80% training set  $(y^{\text{train}}, v^{\text{train}})$  of size  $d_{\text{train}} = 546$ , and a 20% test set  $(y^{\text{test}}, v^{\text{test}})$  of size  $d_{\text{test}} = 137$ , and use Algorithm 7 to draw samples from the predictive distribution  $p(y^{\text{test}}|y^{\text{train}}, v^{\text{train}}, v^{\text{test}}, \bar{\theta}_N)$ . More precisely, we use Algorithm 7 to simultaneously calculate  $\bar{\theta}_N$  and obtain samples from  $p(\beta|y^{\text{train}}, v^{\text{train}}, \bar{\theta}_N)$ , which in turn enables us to simulate from  $p(y^{\text{test}}|\beta, y^{\text{train}}, v^{\text{train}}, v^{\text{test}})$ . We then estimate the maximum-a-posteriori predictive response  $\hat{y}^{\text{test}}$ , and measure prediction accuracy against the test dataset by computing the error

$$\epsilon = \|y^{\text{test}} - \hat{y}^{\text{test}}\|_1 / d_{\text{test}} = \sum_{i=1}^{d_{\text{test}}} |y_i^{\text{test}} - \hat{y}_i^{\text{test}}| / d_{\text{test}}, \quad (5.9)$$

and obtain  $\epsilon = 2.2\%$ .

For comparison, Figure 5.4 below reports the error  $\epsilon$  as a function of  $\theta$  (the discontinuities arise because of the highly non-linear nature of the model). Observe that the estimated  $\bar{\theta}_N$  produces a model that has a very good performance in this regard.



**Figure 5.4** – Bayesian logistic regression - Percentage of mislabelled binary observations in terms of  $\theta$ . In blue we show the value of  $\bar{\theta}_N$  obtained with Algorithm 7.

### 5.3 Audio compressed sensing

Compressed sensing techniques exploit sparsity properties in the data to estimate signals from fewer samples than required by the Nyquist–Shannon sampling theorem [27, 29]. Many real-world data admit a sparse representation on some basis or dictionary. Formally, consider an  $\ell$ -dimensional time-discrete signal  $z \in \mathbb{R}^\ell$  that is sparse in some dictionary  $\Psi \in \mathbb{R}^{\ell \times d}$ , i.e, there exists a latent vector  $x \in \mathbb{R}^d$  such that  $z = \Psi x$  and  $\|x\|_0 = \sum_{i=1}^d \mathbb{1}_{\mathbb{R}^*}(x_i) \ll \ell$ . This prior assumption can be modelled by using a Laplace distribution [91]

$$p(x|\theta) \propto e^{-\theta \|x\|_1}. \quad (5.10)$$

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

$$p(y|x) \propto \exp \left( -\|y - \mathbf{M}\Psi x\|_2^2 / (2\sigma^2) \right), \quad (5.11)$$

leading to the posterior distribution

$$p(x|y, \theta) \propto \exp\left(-\|y - \mathbf{M}\Psi x\|_2^2/(2\sigma^2) - \theta \|x\|_1\right). \quad (5.12)$$

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

$$\hat{x}_{\text{MAP}} \in \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \|y - \mathbf{M}\Psi x\|_2^2/2\sigma^2 + \theta \|x\|_1 \right\}, \quad (5.13)$$

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

Following decades of active research, there are now many convex optimisation algorithms that can be used to efficiently solve (5.13), even when  $d$  is very large [36, 99]. However, the selection of the value of  $\theta$  in (5.13) remains a difficult open problem. This parameter controls the degree of sparsity of  $x$  and has a strong impact on estimation performance.

A common heuristic within the compressed sensing community is to set  $\theta_{\text{cs}} = 0.1 \times \|(\mathbf{M}\Psi)^T y\|_\infty / \sigma^2$ , where for any  $z \in \mathbb{R}^\ell$ ,  $\|z\|_\infty = \max_{i \in \{1, \dots, \ell\}} |z_i|$ , as suggested in [81] and [59]; however, better results can arguably be obtained by adopting a statistical approach to estimate  $\theta$ ; for instance, one can use Algorithm 7 to compute the MLE  $\theta_*$ .

To illustrate this approach, we consider the audio experiment proposed in [12] for the “*Mary had a little lamb*” song. The MIDI-generated audio file  $z$  has  $\ell = 319,725$  samples, but we only have access to a noisy observation vector  $y$  with  $p = 456$  random time points of the audio signal, corrupted by additive white Gaussian noise with  $\sigma = 0.015$ . The latent signal  $x$  has dimension  $d = 2,900$  and is related to  $z$  by a dictionary matrix  $\Psi$  whose row vectors correspond to different piano notes lasting a quarter-second long<sup>3</sup>. We used the heuristic  $\theta_{\text{cs}}$  as the initial value for  $\theta$  in our algorithm. To solve the optimisation problem (5.13) we use the Gradient Projection for Sparse Reconstruction (GPSR) algorithm proposed in [59]. We use this solver because it is the one used in the online MATLAB demonstration of [12], however, more modern algorithms could be used as well.

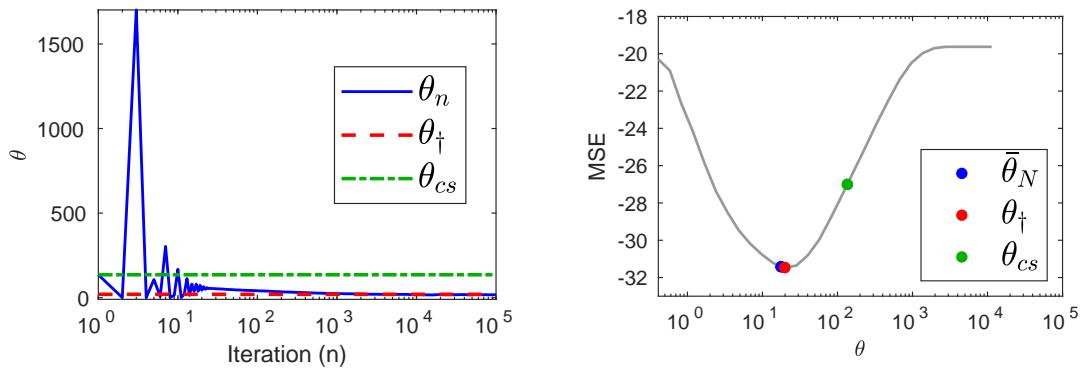
We implemented Algorithm 7 using a step-size  $\gamma = 6.9 \times 10^{-6}$ , a smoothing parameter  $\lambda = 4 \times 10^{-5}$ , a fixed batch size  $m_n = 1$ ,  $\delta_n = 20 n^{-0.8}/d = 0.0069 n^{-0.8}$ ,  $\varphi(\theta) = 0$ , and setting  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 100$  burn-in iterations.

The algorithm converged in approximately 500 iterations, which were computed in only

---

<sup>3</sup>Each quarter-second sound can have one of 100 possible frequencies and be in 29 different positions in time.

325 milliseconds. Figure 5.5 (left), shows the first 250 iterations of the sequence  $\theta_n$  and of the weighted average  $\bar{\theta}_n$ . Again, observe that the iterates oscillate for a few iterations and then quickly stabilise. Finally, to assess the quality of the estimate  $\bar{\theta}_N$ , Figure 5.5 (right) presents the reconstruction mean squared error as a function of  $\theta$ . The error is measured with respect to the reconstructed signal and is given by  $MSE(\hat{x}_{MAP}) = \|z^* - \Psi\hat{x}_{MAP}\|_2^2/\ell$ , where  $z^*$  is the true audio signal. Observe that the estimated value  $\bar{\theta}_N$  is very close to the value that minimises the estimation error, and significantly outperforms the heuristic value  $\theta_{cs}$  commonly used by practitioners.



**Figure 5.5** – Statistical audio compression - Evolution of the iterate  $\theta_n$  with  $\sigma = 0.015$  in log scale (left). Reconstruction mean squared error (MSE) in dB as a function of the  $\theta$  (right). For reference we show the oracle value  $\theta_\dagger$  that minimises the MSE and the frequently used heuristic  $\theta_{cs} = 0.1 \times \|(\mathbf{M}\Psi)^\top y\|_\infty / \sigma^2$ .

Lastly, we want to mention that this approach could be useful for compressed sensing in other domains such as imaging, e.g. [72, 73, 144].

## 5.4 Bayesian logistic regression with random effects

Following on from the Bayesian logistic regression in Section 5.2, where  $p(y|\theta)$  is log-concave and hence  $\theta_*$  unique, we now consider a significantly more challenging sparse Bayesian logistic regression with random effects problem, which is beyond the scope of the theoretical results presented in Appendix C. In this experiment  $p(y|\theta)$  is no longer log-concave, so Algorithm 7 can potentially get trapped in local maximisers. Furthermore, the dimension of  $\theta$  in this experiment is very large ( $d_\theta = 1001$ ), making the MLE problem even more challenging. This experiment was previously considered by [7] and we replicate their setup.

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

$$y_i|x, \beta, \sigma \sim \text{Ber}\left(s(v_i^T \beta + \sigma z_i^T x)\right), \quad i \in \{1, \dots, d_y\}, \quad (5.14)$$

where  $v_i \in \mathbb{R}^p$  are the covariates,  $\beta \in \mathbb{R}^p$  is the regression vector,  $z_i \in \mathbb{R}^d$  are (known) loading vectors,  $x$  are random effects and  $\sigma > 0$ . In addition, recall that  $\text{Ber}(\alpha)$  denotes the Bernoulli distribution with parameter  $\alpha \in [0, 1]$  and  $s(u) = e^u/(1 + e^u)$  is the cumulative distribution function of the standard logistic distribution. The goal is to estimate the unknown parameters  $\theta = (\beta, \sigma) \in \mathbb{R}^p \times (0, +\infty)$  directly from  $\{y_i\}_{i=1}^{d_y}$ , without knowing the value of  $x$ , which we assume to follow a standard Gaussian distribution, *i.e.*  $p(x) = \exp\{-\|x\|_2^2/2\}/(2\pi)^{d/2}$ . We estimate  $\theta$  by MLE using Algorithm 7 to maximise (5.2), with marginal likelihood given by

$$p(y|\theta) = p(y|(\beta, \sigma)) = \int_{\mathbb{R}^d} \prod_{i=1}^{d_y} s(v_i^T \beta + \sigma z_i^T x)^{y_i} (1 - s(v_i^T \beta + \sigma z_i^T x))^{1-y_i} p(x) dx, \quad (5.15)$$

and we use the penalty function

$$\varphi(\theta) = \Lambda \sum_{j=1}^d h_{\Lambda\delta}(\beta_j), \quad (5.16)$$

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

$$h_{\Lambda\delta}(u) = \begin{cases} u^2/(\Lambda\delta 2) & \text{if } |u| \leq \Lambda\delta \\ (|u| - \Lambda\delta/2) & \text{otherwise.} \end{cases}, \quad (5.17)$$

We follow the procedure described in [7] to generate the observations  $\{y_i\}_{i=1}^{d_y}$ , with<sup>4</sup>  $d_y = 500$ ,  $p = 1000$  and  $d = 5$ . The vector of regressors  $\beta_{\text{true}}$  is generated from the uniform distribution on  $[1, 5]$  and 98% of its coefficients are randomly set to zero. The variance  $\sigma_{\text{true}}$  of the random effect is set to 0.1, and the projection interval for the estimated  $\sigma$  is  $[10^{-5}, +\infty)$ . Finally, the parameter  $\Lambda$  in (5.16) is set to  $\Lambda = 30$ . We emphasise at this point that  $\theta$  is high-dimensional in this experiment ( $d_\Theta = 1001$ ), making the estimation problem particularly challenging.

---

<sup>4</sup>We renamed some symbols for notation consistency. What we denote by  $v_i$ ,  $x$ ,  $d_y$  and  $d$ , is denoted in [7] by  $x_i$ ,  $\mathbf{U}$ ,  $N$  and  $q$  respectively.

The conditional log-likelihood function for this model is

$$\log p(y|x, \theta) = \sum_{i=1}^{d_y} \left\{ y_i(v_i^T \beta + \sigma z_i^T x) - \log(1 + e^{v_i^T \beta + \sigma z_i^T x}) \right\}. \quad (5.18)$$

To implement Algorithm 7 we use the gradients

$$\nabla_x \log p(x|y, \theta) = \sum_{i=1}^{d_y} \left\{ \sigma z_i (y_i - s(v_i^T \beta + \sigma z_i^T x)) \right\} - x, \quad (5.19)$$

$$\nabla_\theta \log p(x, y|\theta) = \sum_{i=1}^{d_y} \left\{ (y_i - s(v_i^T \beta + \sigma z_i^T x)) \begin{bmatrix} v_i \\ z_i^T x \end{bmatrix} \right\}. \quad (5.20)$$

Finally, the gradient of the penalty function is given by

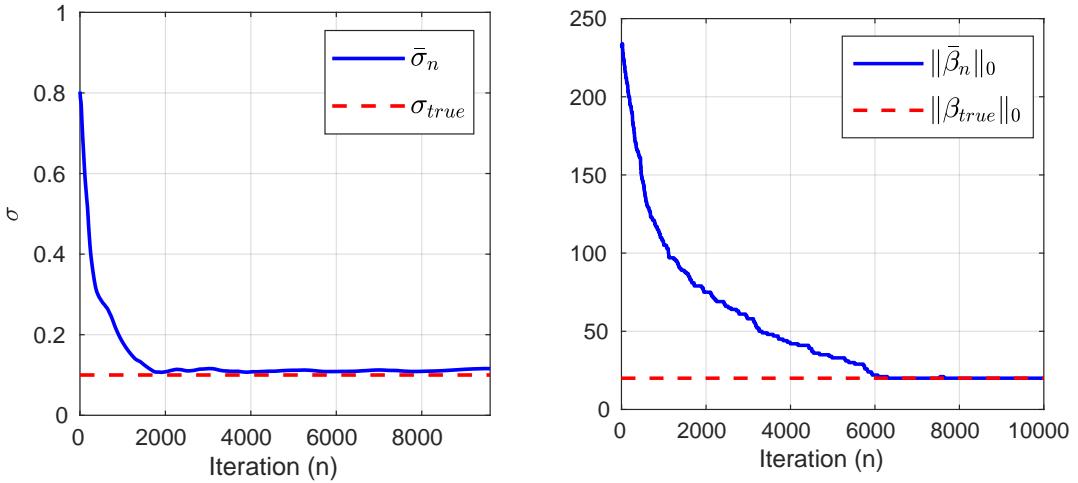
$$\frac{\partial}{\partial \beta_i} \varphi(\theta) = \begin{cases} \beta_i/\delta & |\beta_i| \leq \Lambda\delta \\ \Lambda \operatorname{sign}(\beta_i), & |\beta_i| > \Lambda\delta \end{cases}, \quad \frac{\partial}{\partial \sigma} \varphi(\theta) = 0, \quad (5.21)$$

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

We implement Algorithm 7 using  $\gamma = 0.01$ ,  $\delta_n = n^{-0.95}/d = 0.2 \times n^{-0.95}$ , a fixed batch size  $m_n = 1$ , and with  $\beta_0 = \mathbf{1}_p$  and  $\sigma_0 = 1$  as initial values. We also set the step-size  $\delta$  in (5.16) equal to the step-size  $\delta_n$  in Algorithm 7 to be consistent with [7]. Moreover, we perform  $10^4$  burn-in iterations with a fixed value of  $\theta_0 = (\beta_0, \sigma_0)$  to warm-up the Markov chain, and we set  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 600$  burn-in iterations. Following on from this, we run  $N = 5 \times 10^4$  iterations of Algorithm 7 to compute  $\bar{\theta}_N$ . Computing these estimates required 25 seconds in total.

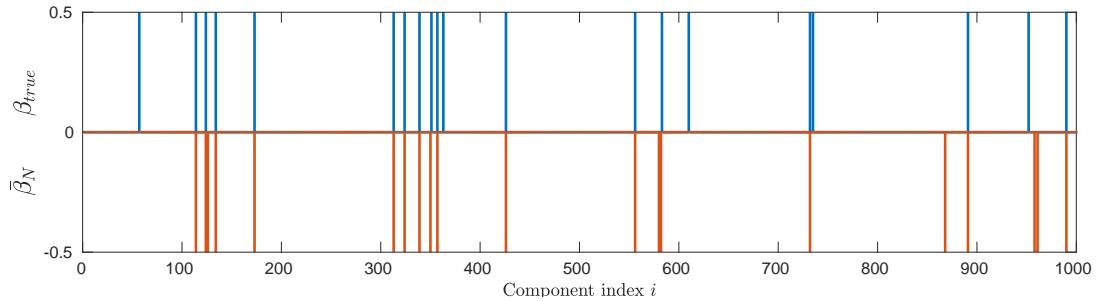
Figure 5.6 shows the evolution of the iterates throughout iterations, where we used  $\|\hat{\beta}_n\|_0$  as a summary statistic to track the number of active components. Because the Huber penalty (5.17) does not enforce exact sparsity on  $\beta$ , to estimate the number of active components we only consider values that are larger than a threshold  $\tau$  (we used  $\tau = 0.005$ ).

From Figure 5.6 we observe that  $\hat{\sigma}_n$  converges to a value that is very close to  $\sigma_{\text{true}}$ , and that the number of active components is also accurately estimated. Moreover, Figure 5.7 shows that most active components were correctly identified. We also observe that  $\hat{\beta}_n$  stabilises after approximately 6300 iterations, which correspond to 6300 Monte Carlo samples as  $m_n=1$ . This is in close agreement with the results presented in [7, Figure



**Figure 5.6** – Sparse Bayesian logistic regression with random effects - Evolution of the  $\|\hat{\beta}_n\|_0$  and of the iterate  $\hat{\sigma}_n$  for the proposed method. The true values are plotted in red as a reference.

5], where they observe stabilisation after a similar number of iterations of their highly specialised Polya-Gamma sampler.



**Figure 5.7** – Sparse Bayesian logistic regression with random effects - Support of the estimated  $\hat{\beta}_N$  compared with the support of  $\beta_{true}$ .

It is worth emphasising at this point that [7] considers the non-smooth penalty  $\varphi(\theta) = \Lambda\|\beta\|_1$  instead of (5.16). Consequently, instead of using the gradient of  $\varphi$ , they resort to the so-called proximal operator of  $\varphi$  [36] with parameter  $\delta_n$ .

# Chapter 6

## Model selection

This chapter presents a fast heuristic for comparing Bayesian models to solve high-dimensional inverse problems. We focus once again on problems that are convex w.r.t. the unknown signal and where no ground truth is available. The proposed heuristic is very computationally efficient and does not require the estimation of the model evidence. Instead, the model evidence is used indirectly to set the regularisation parameters that define each competing model by maximum marginal likelihood estimation, followed by a simple likelihood-based or residual-based comparison of the models based on their empirical Bayesian maximum-a-posteriori solutions. The proposed methodology is illustrated with a total-variation image deblurring experiment, where it performs remarkably well.

### 6.1 Introduction

Many signal processing problems require solving a high-dimensional inverse problem that is ill-conditioned or ill-posed. A key element in all reconstruction methods is the mathematical model that relates the observation to the unknown of interest. This model underpins all inferences about the unknown signal and has a dramatic impact on the estimated results. It is therefore essential to develop advanced tools for selecting and comparing alternative mathematical models.

Indeed, the problem of model selection has received a lot of attention over the past decades. Some of the existing methods, such as cross-validation, rely on the availability of ground truth to compare models. Unfortunately, in many modern applications only a single observation is available, and intrinsic model comparisons under no ground truth can be notoriously difficult.

Bayesian inference offers a variety of approaches for comparing models in these settings

(see [79, Section 1]). For instance, one can compare the posterior probabilities of the models or use an information criterion such as the Widely Applicable BIC [140]. The main difficulty arising in these Bayesian approaches is that the posterior distributions employed in the calculations usually involve intractable integrals. Although many works in the literature explore different strategies for approximating these integrals [41, 61] the resulting methods tend to be either too problem-specific [71] or very computationally expensive (for instance when computing stochastic approximations of the model evidence).

In this chapter we propose a fast heuristic for selecting models without reference to ground truth. The proposed method makes use of the efficient algorithms we proposed earlier in Chapter 3 and is therefore less computationally demanding than other Bayesian approaches. This is mostly because as it does not require estimating the model evidence, but rather uses it indirectly to set the regularisation parameters in each model, and then resorts to a simple residual-based heuristic to compare the competing models.

The remainder of this chapter is organised as follows. Section 6.2 introduces notation and the class of inverse problems considered. Section 6.3 presents the proposed empirical model selection method. In Section 6.4, we illustrate the methodology by selecting one of three possible blur kernels in an image deconvolution problem with a total-variation prior. Conclusions and perspectives are finally reported in Section 6.5.

## 6.2 Bayesian model selection

We are interested in recovering some unknown signal  $x \in \mathbb{R}^n$  from an observation  $y \in \mathbb{R}^m$ . As explained in Section 1.2, we follow a Bayesian approach and specify the posterior distribution  $p(x|y) = p(x, y)/p(y)$  as defined in (1.6). In this chapter, however, we suppose that there are several candidate models  $\mathcal{M}_1, \dots, \mathcal{M}_k$  available to perform inferences on  $x|y$ . Precisely, we suppose that for any  $i \in \{1, \dots, k\}$ ,  $\mathcal{M}_i$  defines a parametric class of log-concave posterior distributions

$$\mathcal{M}_i = \{\theta_i \in \Theta_i : p_i(x|y, \theta_i) = p_i(y|x)p_i(x|\theta_i)/p_i(y|\theta_i)\} \quad (6.1)$$

parametrised by  $\theta_i \in \Theta_i \subset \mathbb{R}^d$ , with likelihood function

$$p_i(y|x) \propto \exp\{-f_{y_i}(x)\}, \quad (6.2)$$

where  $f_{y_i}$  is a convex and  $L_i$ -Lipschitz differentiable function, and with prior distribution given by

$$p_i(x|\theta_i) = \exp\{-\theta_i g_i(x)\}/Z_i(\theta_i), \quad (6.3)$$

where  $g_i$  is a convex, lower-semicontinuous and proper, but possibly non-smooth function. Note that the marginal likelihood  $p_i(y|\theta_i) = \int_{\mathbb{R}^n} p_i(y|x)p_i(x|\theta_i)dx$  is the so-called evidence that measures goodness of fit to  $y$ .

For any specific model  $\mathcal{M}_i$  and value of  $\theta_i$ , a point estimation of  $x|y$  can be readily obtained using proximal optimisation or sampling algorithms to compute the MAP solution given by

$$\hat{x}_i(\theta_i) \in \operatorname{argmin}_{x \in \mathbb{R}^d} f_{y_i}(x) + \theta_i g_i(x). \quad (6.4)$$

This chapter is concerned with the efficient and objective selection of the best Bayesian model  $\mathcal{M}_i$  and regularisation parameter  $\theta_i$  to recover  $x$  from  $y$ . From Bayesian decision theory, one should assign a suitable prior to  $\theta_i$  and select the model with the largest marginal evidence  $p_i(y) = \int_{\Theta_i} p_i(y|\theta_i)p(\theta_i)d\theta_i$  [41, 61]. Alternatively, the empirical Bayesian paradigm proceeds by marginal maximum likelihood estimation and selects the values  $(i, \theta_i) \in \{1, \dots, k\} \times \Theta_i$  that maximise  $p_i(y|\theta_i)$ . Unfortunately, both approaches require the estimation of the evidence  $p_i(y|\theta_i)$ , which is notoriously difficult in imaging problems because of the high dimensionality involved.

We focus instead on a simple heuristic to perform model selection by comparing the likelihoods  $p_i(y|\hat{x}_i(\theta_i^*))$  for  $i \in \{1, \dots, k\}$ , where  $\theta_i^*$  is the marginal likelihood estimator

$$\theta_i^* = \operatorname{argmax}_{\theta_i \in \Theta_i} p_i(y|\theta_i), \quad (6.5)$$

In other words, for each  $i \in \{1, \dots, k\}$ , we identify the value  $\theta_i^*$  that maximises the model evidence  $p_i(y|\theta_i)$  within  $\mathcal{M}_i$ , compute the corresponding (empirical Bayesian) MAP solution  $\hat{x}_i(\theta_i^*)$ , and then compare these based on the likelihoods  $p_i(y|\hat{x}_i(\theta_i^*))$ . This last step is equivalent to comparing the residuals  $r_i = \|y - A_i \hat{x}_i(\theta)\|_2^2$  when  $f_{y_i}(x) = \|y - Ax\|_2^2/2\sigma^2$ . Notice that this heuristic is only as sensible as the strategy to choose  $\theta_i$ , otherwise one would select models that have a small  $r_i$  because they overfit the data<sup>1</sup>.

Before presenting the method to calculate  $\theta_i^*$  to implement our proposed heuristic, we want to stress that this model selection procedure does not carry the rigorous statistical

---

<sup>1</sup>It is usually easy to adjust  $\theta$  to obtain a small residual. In particular, removing the regularisation by setting  $\theta = 0$  leads to smaller residuals but very bad reconstructions  $\hat{x}(\theta)$ . Therefore using residuals for model selection is only reasonable if there is a theoretically underpinned strategy for setting the regularisation parameters.

guarantees of Bayesian model selection procedures. We expect that in some situations it will significantly underperform selection models that compute the evidence  $p_i(y|\theta_i)$ . However, our preliminary numerical experiments suggest that for some problems it can provide a good model selection criterion with a very low computational cost. Also notice that our criterion does not assume that  $p_i(y|\theta_i)$  is finite, and hence allows using improper priors such as total-variation.

### 6.3 Proposed model selection method

The first step in the proposed method, is to set the regularisation parameters of the  $k$  models using the empirical Bayesian method introduced in Chapter 3.

Once the values  $\theta_i^*$  for  $i \in \{1, \dots, k\}$  are estimated by using Algo. 1 or Algo. 3, the proposed model selection heuristic is straight-forward: for each model we obtain the MAP estimator  $\hat{x}_i(\theta_i^*)$  by solving (6.4) with a suitable convex optimisation algorithm, and then calculate the likelihood  $p_i(y|\hat{x}_i(\theta_i^*))$  (or residuals  $r_i = \|y - A_i \hat{x}_i(\theta_i^*)\|_2^2$ ) to compare the competing models. The heuristic is summarised in Algo. 8 below:

---

**Algorithm 8** Empirical model selection
 

---

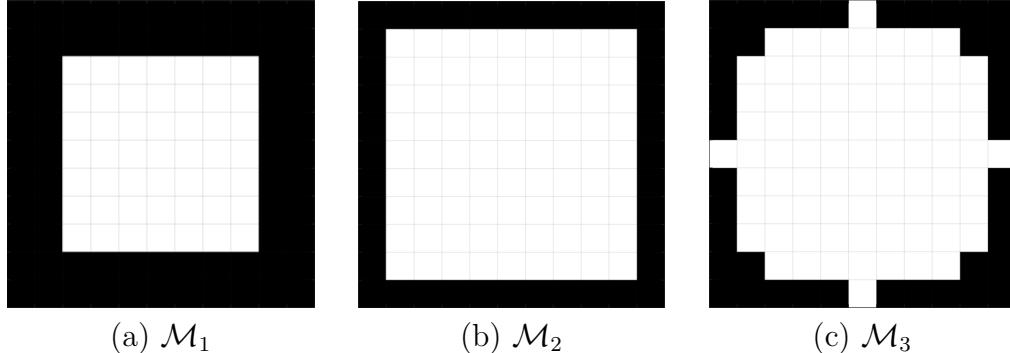
- 1: **for**  $i \in \{1, \dots, k\}$  **do**
  - 2:     Calculate  $\theta_i^*$  using Algorithm 1, Algorithm 2 or Algorithm 3.
  - 3:     Calculate  $\hat{x}_i(\theta_i^*)$  solving (6.4) by convex optimisation.
  - 4:     Calculate the likelihoods  $p_i(y|\hat{x}_i(\theta_i^*))$  (or residual  $r_i$ ).
  - 5: **end for**
  - 6: Select model with the largest likelihood  $p_i(y|\hat{x}_i(\theta_i^*))$  (or smallest residual  $r_i$ )
- 

### 6.4 Numerical experiments

We now illustrate the proposed methodology with a non-blind image deblurring problem for which a series of candidate blur operators are compared to select the best one for restoring a degraded image. We use a total-variation (improper) prior.

More specifically, we want to recover an unknown image  $x \in \mathbb{R}^d$  from a blurred and noisy observation  $y = \mathcal{A}_i x + w$ , where  $\mathcal{A}_i$  is a circulant blurring matrix, and  $w \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{d_y})$ . We consider three candidate statistical models  $\mathcal{M}_1$ ,  $\mathcal{M}_2$  and  $\mathcal{M}_3$ , each defining a posterior distribution (6.1) with  $f_{y_i}(x) = \|y - \mathcal{A}_i x\|_2^2 / 2\sigma^2$ , and  $g_i(x) = TV(x)$  (the isotropic total-variation pseudo-norm) for  $i \in \{1, 2, 3\}$ . Here  $\mathcal{A}_1$  implements a uniform blur of size  $7 \times 7$  pixels,  $\mathcal{A}_2$  implements a uniform blur of size  $9 \times 9$  pixels and

$\mathcal{A}_3$  a circular uniform blur with a 10 pixel diameter. Figure 6.1 shows the point spread functions associated with these blur operators.



**Figure 6.1** – Deblurring with TV prior - Point spread functions for each competing model.

To test the proposed model selection method we proceed as follows. We consider ten standard images of  $512 \times 512$  pixels (`barbara`, `boat`, `bridge`, `flintstones`, `hill`, `lake`, `lena`, `man`, `mandrill` and `wheel`), and for each image we generate 9 different observations  $y$  by using the three different blur operators and three different blurred-signal-to-noise-ratios (SNR) (20 dB, 30 dB and 40 dB). This leads to a total of 90 blurred and noisy test images.

Next, we use the proposed method, as specified in Algorithm 8, to select a model  $\mathcal{M}_i$  for every test image. That is, for each test image, we run Algorithm 1 three times to compute  $\theta_i^*$  for  $i \in \{1, 2, 3\}$ . Then, for each model, we compute the MAP estimator  $\hat{x}_i(\theta_i^*)$  given by (6.4) using the solver SALSA [2], which is an instance of Alternative Direction Method of Multipliers (ADMM). Finally, we compute the corresponding residuals  $r_i$  and select the model with the smallest residual.

To assess the performance of the proposed method, we compare the selected model to the model that gives the best reconstruction mean-squared-error (MSE), which in this problem coincides with the true model. Table 6.1 summarises the results for the 90 test images, organised by SNR value.

We observe from Table 6.1 that the proposed method performs extremely well for medium and high SNR values, where it succeeds at identifying the correct model, in almost every case. Conversely, performance is relatively poor at low SNR values, probably because the posterior distribution in that regime is less concentrated around its mode and the likelihood  $p_i(y|\hat{x}_i(\theta_i^*))$  is a poor surrogate for the marginal

**Table 6.1** – Confusion matrices obtained using 30 different observations (10 images  $\times$  3 kernels) for each SNR value and with a TV prior. The best model is the one that leads to the smallest MSE.

selected			selected			selected					
	$M_1$	$M_2$	$M_3$		$M_1$	$M_2$	$M_3$		$M_1$	$M_2$	$M_3$
true model	$M_1$	10	0	0	$M_1$	10	0	0	$M_1$	10	0
	$M_2$	0	10	0	$M_2$	0	9	1	$M_2$	5	5
	$M_3$	0	0	10	$M_3$	0	0	10	$M_3$	7	0

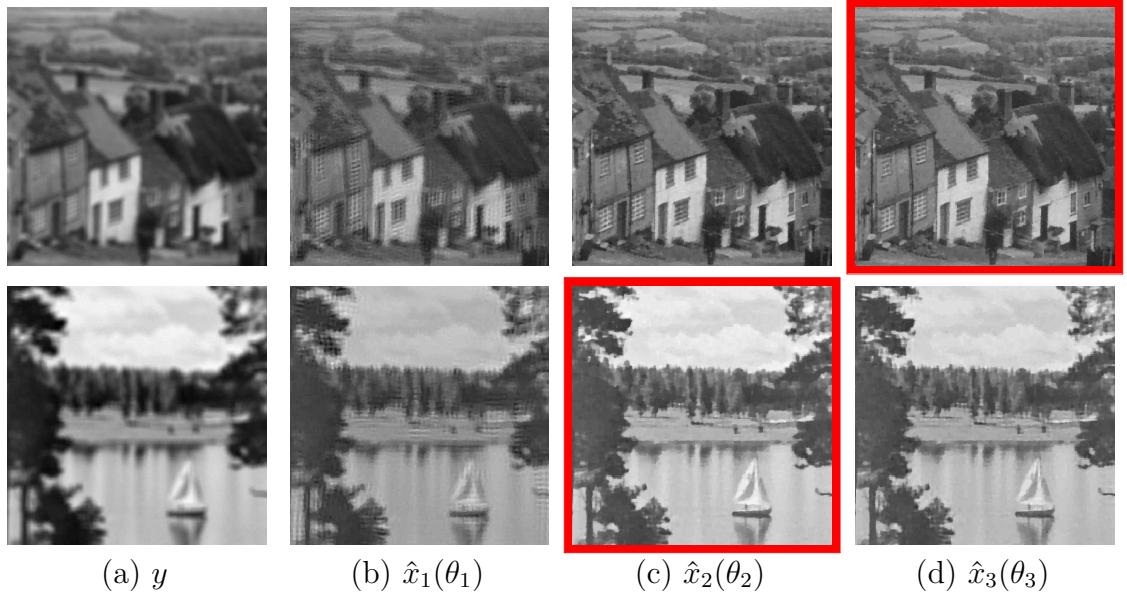
(a) SNR=40dB

(b) SNR=30dB

(c) SNR=20dB

likelihood  $p_i(y|\theta_i^*)$  as a result.

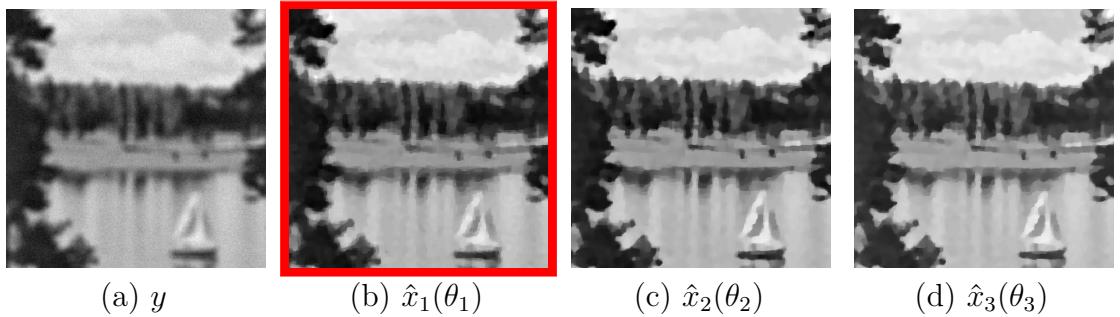
For illustration, Figure 6.2 shows the results obtained for the `hill` and `lake` images for the 30 dB SNR setup, when the best models are  $\mathcal{M}_3$  and  $\mathcal{M}_2$ , respectively. As it may be seen, the algorithm succeeded at selecting the best model, which is remarkable given the proximity between  $\hat{x}_2(\theta_2)$  and  $\hat{x}_3(\theta_3)$  in both cases.



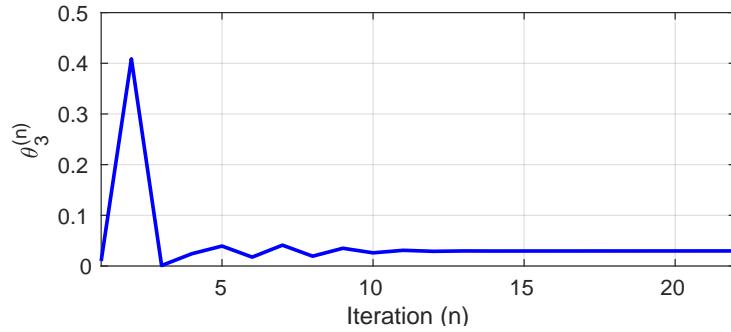
**Figure 6.2** – Deblurring with TV prior - Close-up on `hill` and `lake`: (a) Blurred and noisy image  $y$  (using  $\mathcal{M}_3$  for `hill` and `lake`, and SNR=30 dB), (b)-(d) MAP estimators for  $\mathcal{M}_1$ ,  $\mathcal{M}_2$  and  $\mathcal{M}_3$ . Selected model in red.

In contrast, Figure 6.3 shows the results obtained for the `lake` image for the 20 dB SNR setup where  $\mathcal{M}_3$  is the true model, and in this case the proposed method fails to select the best model (see Figure 6.5 for more details about the specific values of the residuals in each case and the resulting MSE in the reconstructions).

In Figure 6.4 we provide further details for the experiment with SNR=30 dB, where we show the evolution of  $\theta_3^{(n)}$ .



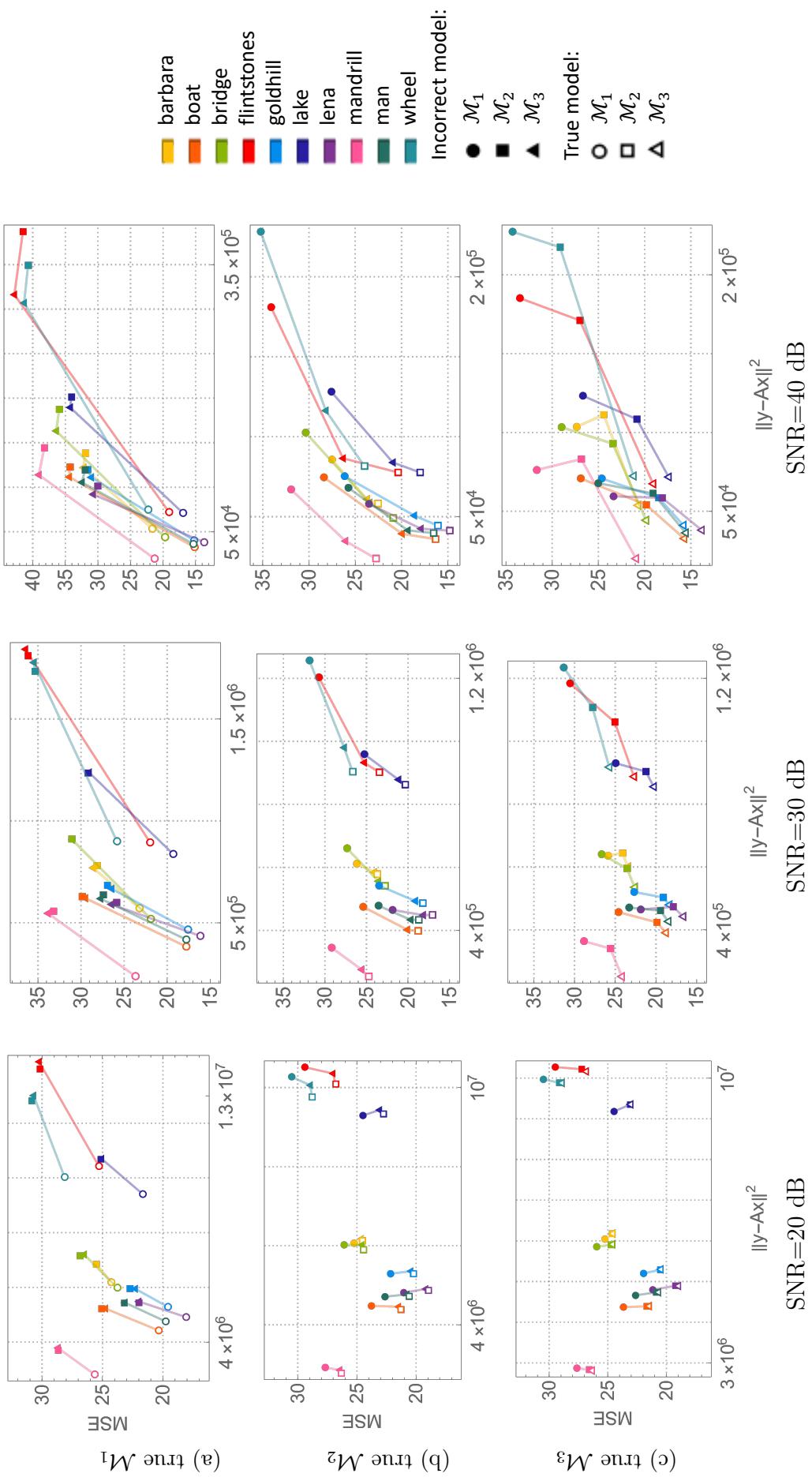
**Figure 6.3** – Deblurring with TV prior - Close-up on `lake`: (a) Blurred and noisy image  $y$  (using  $\mathcal{M}_3$  and SNR=20 dB), (b)-(d) MAP estimators for  $\mathcal{M}_1$ ,  $\mathcal{M}_2$  and  $\mathcal{M}_3$ . Selected model in red (example of a case where it failed to select the true model).



**Figure 6.4** – Deblurring with TV prior on `hill`: evolution of the iterate  $\theta_3^{(n)}$  for Algorithm 1 with SNR=30 dB and true model  $\mathcal{M}_3$ .

To further explore the usefulness of this method for model ranking (and not only model selection) we show in Figure 6.5 nine different scatter plots of MSE against residual  $r_i = \|y - A_i \hat{x}_i(\theta_i^*)\|_2^2$ . The plots are organised in a grid where each column corresponds to a different SNR value and each row corresponds to a different true model (used to generate the observations). Each plot contains 30 points corresponding to 30 different reconstructions  $\hat{x}_i(\theta_i^*)$  (10 observations  $\times$  3 candidate models). The reconstructions corresponding to the same observation are connected with a line for visual clarity.

In this experiment,  $\mathcal{M}_2$  and  $\mathcal{M}_3$  are very similar while  $\mathcal{M}_1$  is further away from both other models. This can be seen in the plots, where squares and triangles (corresponding to reconstructions obtained with  $\mathcal{M}_2$  and  $\mathcal{M}_3$ , respectively) are generally close.



**Figure 6.5 – Deblurring with TV prior -** Scatter plots of residual  $r_i = \|y - A_i \hat{x}_i(\theta_i^*)\|_2^2$  against reconstruction MSE: rows (a), (b) and (c) correspond to cases where the observation  $y$  was generated using  $\mathcal{M}_1$ ,  $\mathcal{M}_2$  and  $\mathcal{M}_3$ , respectively; each column corresponds to a different SNR value. In each plot we show the results for 10 observations (each in a different colour). For each observation, we show the results for the reconstruction  $\hat{x}_i(\theta_i^*)$  obtained with the true model (indicated with an open plot marker) and with the other two candidate models. For visual clarity, we have connected these three results that correspond to the same observation with a line.

A careful inspection of the plots reveals that this approach is not always useful for model ranking and that it depends on the particular models being compared. For example, in the plot corresponding to SNR=40 dB and true  $\mathcal{M}_2$ , the reconstructions with larger residuals also have larger MSE values. In contrast, in the plot corresponding to SNR=40 dB and true  $\mathcal{M}_1$ , the residuals are useful for identifying the best model but fail to rank the relative quality of  $\mathcal{M}_2$  and  $\mathcal{M}_3$ . Lastly, in some cases the behaviour is mixed and depends on the particular observations, e.g. in the plot corresponding to SNR=40 dB and true  $\mathcal{M}_3$  the residuals only fail to rank  $\mathcal{M}_1$  and  $\mathcal{M}_2$  for the observations corresponding to `barbara`, `lena` and `mandril`. Overall, we can conclude that our method is useful for identifying the best model for higher SNR values, but not always for ranking the quality of the suboptimal models. Further research is required to understand what causes the method to fail or succeed at ranking suboptimal models.

Finally, note that we implemented Algorithm 1 following the guidelines in Section 3.3.1 with  $\Theta = [0.001, 1]$ ,  $\theta_0 = 0.01$ , sequence of step-sizes  $\delta_n = 0.1 n^{-0.8}/d$ , and setting  $(\omega_n)_{n \in \mathbb{N}}$  to have  $N_0 = 20$  burn-in iterations. We also start by running 300 warm-up iterations with fixed  $\theta^{(n)} = \theta^{(0)}$ . For the MYULA kernel, we selected  $\gamma$  and  $\lambda$  based on the Lipschitz constant  $L_y$  (which depends on  $\sigma$ ), setting  $\lambda = \min(5L_y^{-1}, \lambda_{\max})$  with  $\lambda_{\max} = 2$  and  $L_y = (0.99/\sigma)^2$ , and  $\gamma = 0.98/(L_y + (1/\lambda))$ . In every case, we stop the algorithm when the relative change in the average value of the iterates  $\theta_i^{(n)}$  is smaller than  $10^{-3}$ . The average computing time for Algorithm 1 was 80 seconds per model <sup>2</sup>.

## 6.5 Conclusions

This chapter presented a computationally efficient method to objectively compare Bayesian models to solve inverse problems related to signal processing, and with a focus on problems that are convex w.r.t. the unknown signal of interest. The method proceeds by first setting the regularisation parameters for each competing model by marginal maximum likelihood estimation by using a proximal MCMC SAPG algorithm. Then, the MAP estimators for each model are retrieved, and the

---

<sup>2</sup>Intel i9-8950HK@2.90GHz workstation running MATLAB R2018a.

resulting likelihoods or residuals w.r.t. the MAP solutions are used as goodness-of-fit measure. The proposed heuristic was illustrated with an application to image deconvolution, where it achieved excellent results in high and medium SNR levels, and more poorly in the low SNR regime.

Perspectives for future work include refining and formalising the proposed heuristic from the lens of the Laplace approximation method<sup>3</sup>, as well as performing additional numerical experiments to assess its performance in other image restoration tasks such as denoising and deblurring with different kinds of regularisers, inpainting, and myopic deconvolution.

---

<sup>3</sup>Recalling that  $p_i(y|\theta_i)$  is given by the prior expectation of the likelihood, then the use of  $p_i(y|\hat{x}_i(\theta_i^*))$  as a surrogate for  $p_i(y|\theta_i)$  could be analysed in terms of a Laplace approximation [8].

# Chapter 7

## Conclusions and perspectives for future work

This thesis 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.

In Chapter 3, we proposed a new computational method to efficiently and accurately estimate regularisation parameters by maximum marginal likelihood estimation, adopting an empirical Bayesian approach. The considered marginal likelihood function is computationally intractable and we addressed this difficulty by using a stochastic proximal gradient optimisation algorithm that is driven by proximal MCMC samplers, and which tightly combines the strengths of modern high-dimensional optimisation and Monte Carlo sampling techniques. Furthermore, we presented three different versions of the algorithm depending on the properties of the regulariser and the tractability of the partition function. We also included a synthetic image denoising problem to study the role of each algorithm parameters and we provided detailed implementation guidelines.

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, it 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. In addition to being highly computationally efficient, the proposed methodology has a strong theoretical underpinning and easily verifiable conditions for convergence (a detailed theoretical analysis of the method is provided in Appendix C).

In Chapter 4, 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 also showed that it is possible to estimate the noise variance jointly with  $\theta$  at a small additional computational cost and with no significant loss of performance in terms of MSE. 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.

In Chapter 5 we showed that the proposed method is also useful for a broader range of intractable maximum likelihood estimation problems. We successfully applied a generalised version of the proposed method to an audio compressive sensing problem, and to two Bayesian logistic regressions with and without random effects. In one of these experiments, we estimated the target marginal log-likelihood and verified that the proposed method yields a very good estimate of its maximiser. We also showed that the method can perform well for estimating more high-dimensional parameters as is the case with one of the logistic regressions which involved an unknown parameter with 1000 components.

In Chapter 6, we explored an application of the proposed methodology to Bayesian model selection. We presented a computationally efficient method to objectively compare Bayesian models which proceeds by i) setting the regularisation parameters for each competing model using the algorithm we propose in Chapter 3, ii) computing the MAP estimators for each model, and iii) using the resulting likelihoods or residuals w.r.t. the MAP solutions as a goodness-of-fit measure. The pro-

posed heuristic was illustrated with an application to image deconvolution, where it achieved excellent results in high and medium SNR levels, and more poorly in the low SNR regime.

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 [134]. 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, as well as to problems involving space-varying regularisation parameters [86]. Lastly, perspectives for future work also include refining and formalising the heuristic for model selection from the lens of the Laplace approximation method, as well as performing additional numerical experiments to assess its performance in other image restoration tasks.

# Appendix A

## Fisher's identity

Fisher's identity is a standard result in the probability literature (e.g. see [54, Proposition D.4]). We reproduce its proof here for completeness.

**Proposition 1.** *For any  $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$  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  $\theta \in \text{int}(\Theta)$ ,  $\theta \mapsto p(y, x|\theta)$  is differentiable. In addition, assume that for any  $y \in \mathbb{R}^{d_y}$  and  $\theta \in \text{int}(\Theta)$ , there exist  $\varepsilon > 0$  and  $\tilde{g}$  such that for any  $\tilde{\theta} \in \overline{B}(\theta, \varepsilon)$  and  $x \in \mathbb{R}^d$ ,  $\|\nabla_\theta p(y, x|\tilde{\theta})\| \leq \tilde{g}(x)$  with  $\int_{\mathbb{R}^d} \tilde{g}(x)p(y|x)dx < +\infty$ . Then, for any  $y \in \mathbb{R}^{d_y}$ ,  $\theta \mapsto p(y|\theta)$  is differentiable over  $\text{int}(\Theta)$  and we have for any  $y \in \mathbb{R}^{d_y}$  and  $\theta \in \text{int}(\Theta)$ ,*

$$\nabla_\theta \log p(y|\theta) = \int_{\mathbb{R}^d} p(x|y, \theta) \nabla_\theta \log p(y, x|\theta) dx . \quad (\text{A.1})$$

*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  $\text{int}(\Theta)$  and we have for any  $\theta \in \text{int}(\Theta)$

$$\nabla_\theta \log p(y|\theta) = \int_{\mathbb{R}^d} p(y|x) \nabla_\theta p(y, x|\theta) dx / p(y|\theta) \quad (\text{A.2})$$

$$= \int_{\mathbb{R}^d} p(y, x|\theta) \nabla_\theta \log p(y, x|\theta) dx / p(y|\theta) = \int_{\mathbb{R}^d} p(x|y, \theta) \nabla_\theta \log p(y, x|\theta) dx , \quad (\text{A.3})$$

which concludes the proof.  $\square$

## Appendix B

### Fair comparison of different methods for setting $\theta$

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  $\theta$  for a given solver, with a given setup (number of iterations, parameters, etc.). Other algorithms, such as the hierarchical one proposed in [105] depend on the solver, but do not seek to optimise  $\theta$  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 [2], which is an efficient implementation of the alternating direction method of multipliers (ADMM) [36]. 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  $MSE(\theta)$  curves for Figure 4.10 (by sampling many points and interpolating), we use SALSA with tolerance  $10^{-5}$  and 1000 iterations as there were some pathological values of  $\theta$  for which SALSA did not converge well with tolerance  $10^{-3}$ . As it may be seen on Figure B.1, 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.

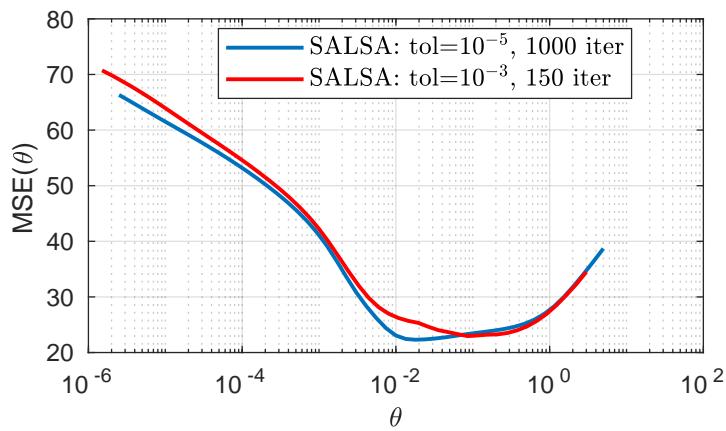
The criterion we opted for was to use SALSA with the strictest tolerance and highest number of iterations, because this configuration gives the overall best estimations.

## B.1 Comparing with solver-dependent methods

As mentioned previously, algorithms like SUGAR try to find the best value of  $\theta$  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  $MSE(\theta)$  curves as the ones in Figure B.1, the values of  $\theta$  computed with SUGAR might yield bad results according to the  $MSE(\theta)$  curve but good results with the specific solver used in SUGAR. For this reason, to achieve a fairer comparison, we compute an equivalent  $\theta_{EQ}$  in the following way. The SUGAR algorithm returns an estimated  $\theta_{SUG}$  and a corresponding  $MSE_{SUG}$  obtained with that  $\theta_{SUG}$ . Given an  $MSE(\theta)$  curve, we define the equivalent  $\theta_{EQ}$  as

$$\theta_{EQ} = \operatorname{argmin}_{\theta \in \Theta} |\theta - \theta_{SUG}| \quad s.t. \quad MSE(\theta) = MSE_{SUG}. \quad (\text{B.1})$$

This  $\theta_{EQ}$  is what we plot in Figure 4.3. For the lowest SNR value,  $\theta_{EQ}$  and  $\theta_{SUG}$  did not differ much in our experiments. However, for the other SNR values, the values of  $\theta_{SUG}$  were significantly smaller than  $\theta_{EQ}$ . This might be related to the fact that SUGAR performed best for the lowest SNR setup.



**Figure B.1** –  $\text{MSE}(\theta)$  for wavelet synthesis- $\ell_1$  deconvolution for SNR = 20dB with boat test image. The curves are computed with different tolerance and maximum iterations using SALSA solver.

## **Appendix C**

### **Analysis of the convergence properties**

In this appendix we include our article [47] that contains all the theoretical proofs of convergence for the proposed methodology. These proofs have been peer reviewed and will soon appear in the SIAM Journal on Imaging Sciences.

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

## Part II: Theoretical Analysis

Valentin De Bortoli <sup>\*1</sup>, Alain Durmus <sup>†1</sup>, Marcelo Pereyra <sup>‡ 2</sup>, and Ana F. Vidal <sup>§</sup>  
 Part of this work has been presented at the 25th IEEE International Conference  
 on Image Processing (ICIP) [50] <sup>2</sup>

<sup>1</sup>CMLA - École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, 94235 Cachan, France.

<sup>2</sup>Maxwell Institute for Mathematical Sciences & School of Mathematical and Computer Sciences,  
 Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom.

August 16, 2020

### Abstract

This paper presents a detailed theoretical analysis of the three stochastic approximation proximal gradient algorithms proposed in our companion paper [49] to set regularization parameters by marginal maximum likelihood estimation. We prove the convergence of a more general stochastic approximation scheme that includes the three algorithms of [49] as special cases. This includes asymptotic and non-asymptotic convergence results with natural and easily verifiable conditions, as well as explicit bounds on the convergence rates. Importantly, the theory is also general in that it can be applied to other intractable optimisation problems. A main novelty of the work is that the stochastic gradient estimates of our scheme are constructed from inexact proximal Markov chain Monte Carlo samplers. This allows the use of samplers that scale efficiently to large problems and for which we have precise theoretical guarantees.

## 1 Introduction

Numerous imaging problems require performing inferences on an unknown image of interest  $x \in \mathbb{R}^d$  from some observed data  $y$ . Canonical examples include image denoising [12, 28], compressive sensing [18, 40], super-resolution [35, 51], tomographic reconstruction [13], image inpainting [24, 44], source separation [9, 8], fusion [46, 31], and phase retrieval [10, 26]. Such imaging problems can be formulated in a Bayesian statistical framework, where inferences are derived from the so-called posterior distribution of  $x$  given  $y$ , which for the purpose of this paper we specify as follows

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

where  $p(y|x) = \exp\{-f_y(x)\}$  with  $f_y \in C^1(\mathbb{R}^d, \mathbb{R})$  is the likelihood function, and the prior distribution is  $p(x|\theta) = \exp\{-\theta^\top g(x)\}$  with  $g : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\Theta}$  and  $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$ . The function  $f_y$  acts as a data-fidelity term,  $g$  as a regulariser that promotes desired structural or regularity properties (e.g., smoothness, piecewise-regularity, or sparsity [11]), and  $\theta$  is a regularisation parameter that controls the amount of regularity enforced. Most Bayesian methods in the imaging literature consider models for which  $f_y$  and  $g$  are convex functions and report as solution the maximum-a-posteriori (MAP) Bayesian estimator

$$\operatorname{argmin}_{x, \theta} f_{y, \theta}, \text{ where } f_{y, \theta}(x) = f_y(x) + \theta^\top g(x) \text{ for any } x \in \mathbb{R}^d. \quad (1)$$

---

<sup>\*</sup>Email: debortoli@cmla.ens-cachan.fr

<sup>†</sup>Email: durmus@cmla.ens-cachan.fr

<sup>‡</sup>Email: m.pereyra@hw.ac.uk

<sup>§</sup>Email: af69@hw.ac.uk

For example, many imaging works consider a linear observation model of the form  $y = Ax + w$ , where  $A \in \mathbb{R}^d \times \mathbb{R}^d$  is some problem-specific linear operator and the noise  $w$  has distribution  $N(0, \sigma^2 \mathbb{I}_d)$  with variance  $\sigma^2 > 0$ . Then, for any  $x \in \mathbb{R}^d$ ,  $f_y(x) = (2\sigma^2)^{-1} \|Ax - y\|^2$ . With regards to the prior, a common choice in imaging is to set  $\Theta = \mathbb{R}^+$  and  $g(x) = \|Bx\|_1$  for some suitable basis or dictionary  $B \in \mathbb{R}^{d'} \times \mathbb{R}^d$ , or  $g(x) = \text{TV}(x)$ , where  $\text{TV}(x)$  is the isotropic total variation pseudo-norm given by  $\text{TV}(x) = \sum_i \sqrt{(\Delta_i^h x)^2 + (\Delta_i^v x)^2}$  where  $\Delta_i^v$  and  $\Delta_i^h$  denote horizontal and vertical first-order local (pixel-wise) difference operators.

Importantly, when  $f_y$  and  $g$  are convex, problem (1) is also convex and can usually be efficiently solved by using modern proximal convex optimisation techniques [11], with remarkable guarantees on the solutions delivered.

Setting the value of  $\theta$  can be notoriously difficult, especially in problems that are ill-posed or ill-conditioned where the regularisation has a dramatic impact on the recovered estimates. We refer to [27] and [49, Section 1] for illustrations and a detailed review of the existing methods for setting set  $\theta$ .

In our companion paper [49], we present a new method to set regularisation parameters. More precisely, in [49], we adopt an empirical Bayesian approach and set  $\theta$  by maximum marginal likelihood estimation, *i.e.*

$$\theta_* \in \arg \max_{\theta \in \Theta} \log p(y|\theta), \quad \text{where } p(y|\theta) = \int_{\mathbb{R}^d} p(y, x|\theta) dx, \quad p(y, x|\theta) \propto \exp[-f_{y,\theta}(x)]. \quad (2)$$

To solve (2), we aim at using gradient based optimization methods. The gradient of  $\theta \mapsto \log p(y|\theta)$ , can be computed using Fisher's identity, see [49, Proposition A.1], which implies under mild integrability conditions on  $f_y$  and  $g$ , for any  $\theta \in \Theta$ ,

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

It follows that  $\theta \mapsto \nabla_\theta \log p(y|\theta)$  can be written as a sum of two parametric integrals which are untractable in most cases. Therefore, we propose to use a stochastic approximation (SA) scheme and, in particular, we define three different algorithms to solve (2) [49, Algorithm 3.1, Algorithm 3.2, Algorithm 3.3]. These algorithms are extensively demonstrated in [49] through a range of applications and comparisons with alternative approaches from the state-of-the-art.

In the present paper we theoretically analyse these three SA schemes and establish natural and easily verifiable conditions for convergence. For generality, rather than presenting algorithm-specific analyses, we establish detailed convergence results for a more general SA scheme that covers the three algorithms of [49] as specific cases. Indeed, all these methods boil down to defining a sequence  $(\theta_n)_{n \in \mathbb{N}}$  satisfying a recursion of the form: for any  $n \in \mathbb{N}$ ,

$$\theta_{n+1} = \Pi_\Theta \left[ \theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \{g(X_k^n) - g(\bar{X}_k^n)\} \right], \quad (3)$$

where  $\Pi_\Theta$  is the projection onto a convex closed set  $\Theta$ ,  $(X_k^n)_{k \in \{1, \dots, m_n\}}$  and  $(\bar{X}_k^n)_{k \in \{1, \dots, m_n\}}$  are two independent stochastic processes targeting  $x \mapsto p(x|y, \theta)$  and  $x \mapsto p(x|\theta)$  respectively,  $(m_n)_{n \in \mathbb{N}}$  is a sequence of batch-sizes and  $(\delta_n)_{n \in \mathbb{N}^*}$  is a sequence of stepsizes. In this paper, we are interested in establishing the convergence of the averaging of  $(\theta_n)_{n \in \mathbb{N}}$  to a solution of (2) in this setting. SA has been extensively studied during the past decades [41, 29, 38, 47, 33, 34, 7, 6, 48]. Recently, quantitative results have been obtained in [45, 2, 39, 1, 43]. In contrast to [1], here we consider the case where  $(X_k^n)_{k \in \{1, \dots, m_n\}}$  and  $(\bar{X}_k^n)_{k \in \{1, \dots, m_n\}}$  are *inexact* Markov chains which target  $x \mapsto p(x|y, \theta)$  and  $x \mapsto p(x|\theta)$  respectively and are based on some generalizations of the Unadjusted Langevin Algorithm (ULA) [42]. In the recent years, ULA has attracted a lot of attention since this algorithm exhibits favorable high-dimensional convergence properties in the case where the target distribution admits a differentiable density, see [20, 22, 14, 15]. However, in most imaging models, the penalty function  $g$  is not differentiable and therefore  $x \mapsto p(x|y, \theta)$  and  $x \mapsto p(x|\theta)$  are not differentiable as well. Therefore, we consider proximal Langevin samplers which are specifically designed to overcome this issue: the Moreau-Yoshida Unadjusted Langevin Algorithm (MYULA), see [23], and the Proximal Unadjusted Langevin Operator (PULA), see [21].

A similar approximation scheme to (3) is studied in [1]. More precisely [1, Theorem 3, Theorem 4] are similar to Theorem 6 and Theorem 7. Contrarily to that work, here we do not require the Markov kernels we use to exactly target  $x \mapsto p(x|\theta)$  and  $x \mapsto p(x|y, \theta)$  but allow some bias in the estimation which is accounted for in our convergence rates. This relaxation to biased

estimates plays a central role in the capacity of the method to scale efficiently to large problems. Moreover, the present paper is also a complement of [17] which establishes general conditions for the convergence of inexact Markovian SA but only apply these results to ULA. In this study, we do not consider a general Markov kernel but rather specialize the results of [17] to MYULA and PULA Markov kernels. However, to apply results of [17], new quantitative geometric convergence properties on MYULA and PULA have to be established.

The remainder of the paper is organized as follows. In Section 2, we recall our notations and conventions. In Section 3, we define the class of optimisation problems considered and the SA scheme (3). This setting includes the optimization problem presented in (2) and the three specific algorithms introduced in [49]. Then, in Section 4, we present a detailed analysis of the theoretical properties of the proposed methodology. First, we show new ergodicity results for the MYULA and PULA samplers. In a second part, we provide easily verifiable conditions for convergence and quantitative convergence rates for the averaging sequences designed from (3). The proofs of these results are gathered in Section 5.

## 2 Notations and conventions

We denote by  $B(0, R)$  and  $\overline{B}(0, R)$  the open ball, respectively the closed ball, with radius  $R$  in  $\mathbb{R}^d$ . Denote by  $\mathcal{B}(\mathbb{R}^d)$  the Borel  $\sigma$ -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  $\mu$  a probability measure on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$  and  $f \in \mathbb{F}(\mathbb{R}^d)$  a  $\mu$ -integrable function, denote by  $\mu(f)$  the integral of  $f$  w.r.t.  $\mu$ . 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  $\xi$  be a finite signed measure on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ . The  $V$ -total variation norm of  $\xi$  is defined as

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

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

Let  $U$  be an open set of  $\mathbb{R}^d$ . We denote by  $C^k(U, \mathbb{R}^{d\Theta})$  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.  $C^k(U)$  stands  $C^k(U, \mathbb{R})$ . Let  $f : U \rightarrow \mathbb{R}$ , we denote by  $\nabla f$ , the gradient of  $f$  if it exists.  $f$  is said to be  $m$ -convex with  $m \geq 0$  if for all  $x, y \in \mathbb{R}^d$  and  $t \in [0, 1]$ ,

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

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

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

## 3 Proposed stochastic approximation proximal gradient optimisation methodology

### 3.1 Problem statement

Let  $\Theta \subset \mathbb{R}^{d\Theta}$  and  $f : \Theta \rightarrow \mathbb{R}$ . We consider the optimisation problem

$$\theta_* \in \arg \min_{\theta \in \Theta} f(\theta) , \quad (4)$$

in scenarios where it is not possible to evaluate  $f$  nor  $\nabla f$  because they are computationally intractable. Problem (4) includes the marginal likelihood estimation problem (2) of our companion paper [49] as the special case  $f = -\log p(y|\cdot)$ . We make the following general assumptions on  $f$  and  $\Theta$ , which are in particular verified by the imaging models considered in [49].

**A1.**  $\Theta$  is a convex compact set and  $\Theta \subset \overline{B}(0, R_\Theta)$  with  $R_\Theta > 0$ .

**A2.** There exist an open set  $U \subset \mathbb{R}^p$  and  $L_f \geq 0$  such that  $\Theta \subset U$ ,  $f \in C^1(U, \mathbb{R})$  and for any  $\theta_1, \theta_2 \in \Theta$

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

**A3.** For any  $\theta \in \Theta$ , there exist  $H_\theta, \bar{H}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\Theta}$  and two probability distributions  $\pi_\theta, \bar{\pi}_\theta$  on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$  satisfying for any  $\theta \in \Theta$

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

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

**Remark 1.** Note that if  $f \in C^2(\Theta)$  then **A2** is automatically satisfied under **A1**, since  $\Theta$  is compact. In every model considered in our companion paper [49],  $\theta \mapsto -\log p(y|\theta)$  is continuously twice differentiable on each compact using the dominated convergence theorem and therefore **A2** holds under **A1**.

**Remark 2.** Assumption **A3** is verified in the three cases considered in our companion paper [49, Algorithm 3.1, Algorithm 3.2, Algorithm 3.3]:

- (a) if the regulariser  $g$  is  $\alpha$  positively homogeneous with  $\alpha > 0$  and  $d_\Theta = 1$ , corresponding to [49, Algorithm 3.1], then for any  $\theta \in \Theta$ ,  $H_\theta = g$ ,  $\bar{H}_\theta = -d/(\alpha\theta)$ ,  $\pi_\theta$  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;
- (b) if the regulariser  $g$  is separably positively homogeneous as in [49, Algorithm 3.2], then for any  $\theta \in \Theta$ ,  $H_\theta = g$ ,  $\bar{H}_\theta = (-|\mathbf{A}_i|/(\alpha_i\theta^i))_{i \in \{1, \dots, d_\Theta\}}$ ,  $\pi_\theta$  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;
- (c) if the regulariser  $g$  is inhomogeneous, corresponding to [49, Algorithm 3.3], then for any  $\theta \in \Theta$ ,  $\bar{H}_\theta = -g$ ,  $H_\theta = g$ ,  $\pi_\theta$  and  $\bar{\pi}_\theta$  are the probability measures associated with the posterior and the prior, with density w.r.t. the Lebesgue measure  $x \mapsto p(x|y, \theta)$  and  $x \mapsto p(x|\theta)$  respectively.

We now present in Algorithm 1, the stochastic algorithm we consider in order to solve (4). This method encompasses the schemes introduced in the companion paper [49, Algorithm 3.1, Algorithm 3.2, Algorithm 3.3]. 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  $k \in \{0, \dots, m_n - 1\}$

$$\begin{aligned} (X_k^n)_{k \in \{0, \dots, m_n\}} &\text{ is a MC with kernel } K_{\gamma_n, \theta_n} \text{ and } X_0^n = X_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1}, \\ (\bar{X}_k^n)_{k \in \{0, \dots, m_n\}} &\text{ is a MC with kernel } \bar{K}_{\gamma'_n, \theta_n} \text{ and } \bar{X}_0^n = \bar{X}_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1}, \\ \theta_{n+1} &= \Pi_\Theta \left[ \theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \{H_{\theta_n}(X_k^n) + \bar{H}_{\theta_n}(\bar{X}_k^n)\} \right], \end{aligned} \quad (5)$$

where  $(X_{m-1}^{-1}, \bar{X}_{m-1}^{-1}) = (X_0^0, \bar{X}_0^0)$ ,  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma > 0, \theta \in \Theta\}$  is a family of Markov kernels on  $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ ,  $(m_n)_{n \in \mathbb{N}} \in (\mathbb{N}^*)^\mathbb{N}$ ,  $\delta_n, \gamma_n, \gamma'_n > 0$  for any  $n \in \mathbb{N}$ ,  $\Pi_\Theta$  is the projection onto  $\Theta$  and  $\mathcal{F}_n$  is defined as follows for all  $n \in \mathbb{N} \cup \{-1\}$

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

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

$$\bar{\theta}_N = \sum_{n=0}^{N-1} \delta_n \theta_n \Bigg/ \sum_{n=0}^{N-1} \delta_n.$$

In the sequel, we are interested in the convergence of  $(f(\bar{\theta}_N))_{N \in \mathbb{N}}$  to a minimum of  $f$  in the case where the Markov kernels  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma > 0, \theta \in \Theta\}$ , used in Algorithm 1 are either the ones associated with MYULA or PULA. We now present these two MCMC methods for which some analysis is required in our study of  $(f(\bar{\theta}_N))_{N \in \mathbb{N}}$ .

### 3.2 Choice of MCMC kernels

Given the high dimensionality involved, it is fundamental to carefully choose the families of Markov kernels  $\{K_{\gamma, \theta}, \bar{K}_{\gamma, \theta} : \gamma > 0, \theta \in \Theta\}$  driving Algorithm 1. In the experimental part of this work, see [49, Section 4], we use the MYULA Markov kernel recently proposed in [23], which is a state-of-the-art proximal Markov chain Monte Carlo (MCMC) method specifically designed for high-dimensional models that are log-concave but not smooth. The method is derived from the

---

**Algorithm 1** General algorithm

---

```

1: Input: initial  $\{\theta_0, X_0^0, \bar{X}_0^0\}$ ,  $(\delta_n, \gamma_n, \gamma'_n, m_n)_{n \in \mathbb{N}}$ , number of iterations  $N$ .
2: for  $n = 0$  to  $N - 1$  do
3:   if  $n > 0$  then
4:     Set  $X_0^n = X_{m_{n-1}}^{n-1}$ ,
5:     Set  $\bar{X}_0^n = \bar{X}_{m_{n-1}}^{n-1}$ ,
6:   end if
7:   for  $k = 0$  to  $m_n - 1$  do
8:     Sample  $X_{k+1}^n \sim K_{\gamma_n, \theta_n}(X_k^n, \cdot)$ ,
9:     Sample  $\bar{X}_{k+1}^n \sim \bar{K}_{\gamma'_n, \theta_n}(\bar{X}_k^n, \cdot)$ ,
10:   end for
11:   Set  $\theta_{n+1} = \Pi_\Theta \left[ \theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} \{H_{\theta_n}(X_k^n) + \bar{H}_{\theta_n}(\bar{X}_k^n)\} \right]$ .
12: end for
13: Output:  $\bar{\theta}_N = \{\sum_{n=0}^{N-1} \delta_n\}^{-1} \sum_{n=0}^{N-1} \delta_n \theta_n$ .

```

---

discretisation of an over-damped Langevin diffusion,  $(\bar{X}_t)_{t \geq 0}$ , satisfying the following stochastic differential equation

$$d\mathbf{X}_t = -\nabla_x F(\mathbf{X}_t)dt + \sqrt{2}dB_t, \quad (6)$$

where  $F : \mathbb{R}^d \mapsto \mathbb{R}$  is a continuously differentiable potential and  $(B_t)_{t \geq 0}$  is a standard  $d$ -dimensional Brownian motion. Under mild assumptions, this equation has a unique strong solution [25, Chapter 4, Theorem 2.3]. Accordingly, the law of  $(X_t)_{t \geq 0}$  converges as  $t \rightarrow \infty$  to the diffusion's unique invariant distribution, with probability density given by  $\pi(x) \propto e^{-F(x)}$  for all  $x \in \mathbb{R}^d$  [42, Theorem 2.2]. Hence, to use (6) as a Monte Carlo method to sample from the posterior  $p(x|y, \theta)$ , we set  $F(x) = \log p(x|y, \theta)$  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 (6) is usually not computationally feasible. Instead, we usually resort to a discrete-time Euler-Maruyama approximation of (6) 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

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

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) [42]. Under some additional assumptions on  $F$ , namely Lipschitz continuity of  $\nabla_x F$ , the ULA chain inherits the convergence properties of (6) and converges to a stationary distribution that is close to the target  $\pi$ , with  $\gamma$  controlling a trade-off between accuracy and convergence speed [23].

**Remark 3.** In this form, the ULA algorithm is limited to distributions where  $F$  is a Lipschitz continuously differentiable function. However, in the imaging problems of interest this is usually not the case [49]. For example, to implement any of the algorithms presented in [49] it is necessary to sample from the posterior distribution  $p(x|y, \theta)$  (corresponding to  $\pi_\theta$  in Section 3.1), which would require setting for any  $x \in \mathbb{R}^d$ ,  $F(x) = f_y(x) + \theta^\top g(x)$ . Similarly, one of the algorithms also requires sampling from the prior distribution  $x \mapsto p(x|\theta)$  (corresponding to  $\bar{\pi}_\theta$  in Section 3.1), which requires setting for any  $x \in \mathbb{R}^d$ ,  $F(x) = \theta^\top g(x)$ . In both cases, if  $g$  is not smooth then ULA cannot be directly applied. The MYULA kernel was designed precisely to overcome this limitation.

### 3.2.1 Moreau-Yoshida Unadjusted Langevin Algorithm

Suppose that the target potential admits a decomposition  $F = V + U$  where  $V$  is Lipschitz differentiable and  $U$  is not smooth but convex over  $\mathbb{R}^d$ . In MYULA, the differentiable part is handled via the gradient  $\nabla_x V$  in a manner akin to ULA, whereas the non-differentiable convex part is replaced by a smooth approximation  $U^\lambda(x)$  given by the Moreau-Yosida envelope of  $U$ , see [5, Definition 12.20], defined for any  $x \in \mathbb{R}^d$  and  $\lambda > 0$  by

$$U^\lambda(x) = \min_{\tilde{x} \in \mathbb{R}^d} \left\{ U(\tilde{x}) + (1/2\lambda) \|x - \tilde{x}\|_2^2 \right\}. \quad (7)$$

Similarly, we define the proximal operator for any  $x \in \mathbb{R}^d$  and  $\lambda > 0$  by

$$\text{prox}_U^\lambda(x) = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left\{ U(\tilde{x}) + (1/2\lambda) \|x - \tilde{x}\|_2^2 \right\}. \quad (8)$$

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

$$\nabla U^\lambda(x) = (x - \text{prox}_U^\lambda(x))/\lambda, \quad (9)$$

(see, e.g., [5, 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  $k \in \mathbb{N}$

$$\text{MYULA} : X_{k+1} = X_k - \gamma \nabla_x V(X_k) - \gamma \nabla_x U^\lambda(X_k) + \sqrt{2\gamma} Z_{k+1}. \quad (10)$$

Returning to the imaging problems of interest, we define the MYULA families of Markov kernels  $\{\mathbf{R}_{\gamma,\theta}, \bar{\mathbf{R}}_{\gamma,\theta} : \gamma > 0, \theta \in \Theta\}$  that we use in Algorithm 1 to target  $\pi_\theta$  and  $\bar{\pi}_\theta$  for  $\theta \in \Theta$  as follows. By Remark 3, we set  $V = f_y$  and  $U = \theta^\top g$ ,  $\bar{V} = 0$  and  $\bar{U} = \theta^\top g$ . Then, for any  $\theta \in \Theta$  and  $\gamma > 0$ ,  $\mathbf{R}_{\gamma,\theta}$  associated with  $(X_k)_{k \in \mathbb{N}}$  is given by  $X_0 \in \mathbb{R}^d$  and the following recursion for any  $k \in \mathbb{N}$

$$X_{k+1} = X_k - \gamma \nabla_x f_y(X_k) - \gamma \left\{ X_k - \text{prox}_{\theta^\top g}^\lambda(X_k) \right\} / \lambda + \sqrt{2\gamma} Z_{k+1}. \quad (11)$$

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

$$\bar{X}_{k+1} = \bar{X}_k - \gamma' \left\{ \bar{X}_k - \text{prox}_{\theta^\top g}^{\lambda'}(\bar{X}_k) \right\} / \lambda' + \sqrt{2\gamma} Z_{k+1}, \quad (12)$$

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.

Notice that other ways of splitting the target potential  $F$  can be straightforwardly implemented. For example, instead of a single non-smooth convex term  $U$ , one might choose a splitting involving several non-smooth terms to simplify the computation of the proximal operators (each term would be replaced by its Moreau-Yosida envelope in (6)). Similarly, although we usually associate  $V, \bar{V}$  and  $U, \bar{U}$  to the log-likelihood and the log-prior, some cases might benefit from a different splitting. Moreover, as illustrated in Section 3.2.2 below, other discrete approximations of the Langevin diffusion could be considered too.

### 3.2.2 Proximal Unadjusted Langevin Algorithm

As an alternative to MYULA, one could also consider using the Proximal Unadjusted Langevin Algorithm (PULA) introduced in [21], which replaces the (forward) gradient step of MYULA by a composition of a backward and forward step. More precisely, PULA defines the Markov chain  $(X_k)_{k \in \mathbb{N}}$  starting from  $X_0 \in \mathbb{R}^d$  by the following recursion: for any  $k \in \mathbb{N}$

$$\text{PULA} : X_{k+1} = \text{prox}_U^\lambda(X_k) - \gamma \nabla_x U(\text{prox}_U^\lambda(X_k)) + \sqrt{2\gamma} Z_{k+1}. \quad (13)$$

To highlight the connection with MYULA we note that for any  $x \in \mathbb{R}^d$  and  $\lambda \geq 0$ ,  $\nabla U^\lambda(x) = (x - \text{prox}_U^\lambda(x))/\lambda$  by [5, Proposition 12.30]. Therefore, if we set  $\lambda = \gamma$  we obtain that (13) can be rewritten for any  $k \in \mathbb{N}$  a

$$X_{k+1} = X_k - \gamma \nabla_x V(X_k) - \gamma \nabla_x U(\text{prox}_U^\lambda(X_k)) + \sqrt{2\gamma} Z_{k+1},$$

which corresponds to (10) with  $\lambda = \gamma$ , except that the term  $\nabla_x U(X_k)$  in (10) is replaced by  $\nabla_x U(\text{prox}_U^\lambda(X_k))$  in (10).

Going back to the imaging problems of interest, to define the PULA families of Markov kernels  $\{\mathbf{S}_{\gamma,\theta}, \bar{\mathbf{S}}_{\gamma,\theta} : \gamma > 0, \theta \in \Theta\}$  that we use in Algorithm 1 to target  $\pi_\theta$  and  $\bar{\pi}_\theta$  for  $\theta \in \Theta$  we proceed as follows. We set  $V = f_y$  and  $U = \theta^\top g$ ,  $\bar{V} = 0$  and  $\bar{U} = \theta^\top g$ . Then, by Remark 3, for any  $\theta \in \Theta$  and  $\gamma > 0$ ,  $\mathbf{S}_{\gamma,\theta}$  associated with  $(X_k)_{k \in \mathbb{N}}$  is given by  $X_0 \in \mathbb{R}^d$  and the following recursion for any  $k \in \mathbb{N}$

$$X_{k+1} = \text{prox}_{\theta^\top g}^\lambda(X_k) - \gamma \nabla_x f_y(\text{prox}_{\theta^\top g}^\lambda(X_k)) + \sqrt{2\gamma} Z_{k+1}, \quad (14)$$

Similarly, for any  $\theta \in \Theta$  and  $\gamma' > 0$ ,  $\bar{\mathbf{S}}_{\gamma,\theta}$  associated with  $(X_k)_{k \in \mathbb{N}}$  is given by  $X_0 \in \mathbb{R}^d$  and the following recursion for any  $k \in \mathbb{N}$

$$\bar{X}_{k+1} = \text{prox}_{\theta^\top g}^{\lambda'}(\bar{X}_k) + \sqrt{2\gamma} Z_{k+1}. \quad (15)$$

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. Again, one could use PULA with a different splitting of  $F$ .

Finally, we note at this point that the MYULA and PULA kernels (11), (12), (14) and (15), 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 (6), 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 1 will exhibit some asymptotic estimation bias. This error is controlled by  $\lambda, \lambda', \gamma, \gamma'$ , and  $\delta$ , and can be made arbitrarily small at the expense of additional computing time, see Theorem 7 in Section 4.

## 4 Analysis of the convergence properties

### 4.1 Ergodicity properties of MYULA and PULA

Before establishing our main convergence results about Algorithm 1, see Section 4.1, we derive ergodicity properties on the Markov chains given by (10) and (13). We consider the following assumptions on  $\pi_\theta$  and  $\bar{\pi}_\theta$ . These assumptions are satisfied for a large class of models in Bayesian imaging sciences, and in particular by the models considered in our companion paper [49].

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

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

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

*and*

$$\min \left( \inf_{\theta \in \Theta} \int_{\mathbb{R}^d} \exp[-V_\theta(\tilde{x}) - U_\theta(\tilde{x})] d\tilde{x}, \inf_{\theta \in \Theta} \int_{\mathbb{R}^d} \exp[-\bar{V}_\theta(\tilde{x}) - \bar{U}_\theta(\tilde{x})] d\tilde{x} \right) > 0 . \quad (16)$$

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

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

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

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

$$\max (\|U_\theta(x) - U_\theta(y)\|, \|\bar{U}_\theta(x) - \bar{U}_\theta(y)\|) \leq M \|x - y\| .$$

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

Note that (16) in H1-(a) is satisfied if  $\Theta$  is compact and the functions  $\theta \mapsto \int_{\mathbb{R}^d} \exp[-V_\theta(\tilde{x}) - U_\theta(\tilde{x})] d\tilde{x}$  and  $\theta \mapsto \int_{\mathbb{R}^d} \exp[-\bar{V}_\theta(\tilde{x}) - \bar{U}_\theta(\tilde{x})] d\tilde{x}$  are continuous. This latter condition can be 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  $V : \mathbb{R}^d \rightarrow [0, +\infty)$  such that for any  $\theta \in \Theta$ ,  $V_\theta = V$  and there exists  $x^* \in \mathbb{R}^d$  with  $x^* \in \arg \min_{\mathbb{R}^d} V$  then one can choose  $x_\theta^* = x^*$  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  $x_\theta^\sharp = 0$  in H1-(c) and in this case  $R_{U,1} = R_{U,2} = 0$ . These conditions are satisfied by all the models studied in [49].

As emphasized in Section 3.1, we use a stochastic approximation proximal gradient approach to minimize  $f$  and therefore we need to consider Monte Carlo estimators for  $\nabla_\theta f(\theta)$  and  $\theta \in \Theta$ . These estimators are derived from Markov chains targeting  $\pi_\theta$  and  $\bar{\pi}_\theta$  respectively. We consider two MCMC methodologies to construct the Markov chains. A first option, as proposed in Section 3.2.1, is to use MYULA to sample from  $\pi_\theta$  and  $\bar{\pi}_\theta$ . Let  $\kappa > 0$  and  $\{R_{\gamma,\theta} : \gamma > 0, \theta \in \Theta\}$  be the family of kernels defined for any  $x \in \mathbb{R}^d$ ,  $\gamma > 0$ ,  $\theta \in \Theta$  and  $A \in \mathcal{B}(\mathbb{R}^d)$  by

$$R_{\gamma,\theta}(x, A) = (4\pi\gamma)^{-d/2} \int_A \exp \left( \frac{\|y - x + \gamma \nabla_x V_\theta(x) + \kappa^{-1} \{x - \text{prox}_{U_\theta}^{\gamma\kappa}(x)\}\|^2}{(4\gamma)} \right) dy . \quad (17)$$

Note that (17) is the Markov kernel associated with the recursion (10) with  $U \leftarrow U_\theta$ ,  $V \leftarrow V_\theta$  and  $\lambda \leftarrow \kappa\gamma$ . For any  $\gamma, \kappa > 0$  and  $\theta \in \Theta$  corresponds to  $R_{\gamma, \kappa\gamma, \theta}$  in [49]. Consider also the family of Markov kernels  $\{\bar{R}_{\gamma, \theta} : \gamma > 0, \theta \in \Theta\}$  such that for any  $\gamma > 0$  and  $\theta \in \Theta$ ,  $\bar{R}_{\gamma, \theta}$  is the Markov kernel defined by (17) but with  $\bar{U}_\theta$  and  $\bar{V}_\theta$  in place of  $U_\theta$  and  $V_\theta$  respectively. The coefficient  $\kappa$  is related to  $\lambda$  in (11) by  $\kappa = \lambda/\gamma$ .

Moreover, although our companion paper [49] only considers the MYULA kernel, the theoretical results we present in this paper also hold if the algorithms are implemented using PULA [21]. Define the family  $\{S_{\gamma, \theta} : \gamma > 0, \theta \in \Theta\}$ , for any  $x \in \mathbb{R}^d$ ,  $\gamma > 0$ ,  $\theta \in \Theta$  and  $A \in \mathcal{B}(\mathbb{R}^d)$  by

$$S_{\gamma, \theta}(x, A) = (4\pi\gamma)^{-d/2} \int_A \exp\left(\|y - \text{prox}_{U_\theta}^{\gamma\kappa}(x) + \gamma\nabla_x V_\theta(\text{prox}_{U_\theta}^{\gamma\kappa}(x))\|^2/(4\gamma)\right) dy. \quad (18)$$

Note that (17) is the Markov kernel associated with the recursion (13) with  $U \leftarrow U_\theta$ ,  $V \leftarrow V_\theta$  and  $\lambda \leftarrow \kappa\gamma$ . Consider also the family of Markov kernels  $\{\bar{S}_{\gamma, \theta} : \gamma > 0, \theta \in \Theta\}$  such that for any  $\gamma > 0$  and  $\theta \in \Theta$ ,  $\bar{S}_{\gamma, \theta}$  is the Markov kernel defined by the recursion (18) but with  $\bar{U}_\theta$  and  $\bar{V}_\theta$  in place of  $U_\theta$  and  $V_\theta$  respectively. We use the results derived in [17] to analyse the sequence given by (5) with  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(R_{\gamma, \theta}, \bar{R}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$  or  $\{(S_{\gamma, \theta}, \bar{S}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ . To this end, we impose that for any  $\gamma \in (0, \bar{\gamma}]$  and  $\theta \in \Theta$ , the kernels  $K_{\gamma, \theta}$  and  $\bar{K}_{\gamma, \theta}$  admit an invariant probability distribution, denoted by  $\pi_{\gamma, \theta}$  and  $\bar{\pi}_{\gamma, \theta}$  respectively which are approximations of  $\pi_\theta$  and  $\bar{\pi}_\theta$  defined in A3, and geometrically converge towards them. More precisely, we show in Theorem 4 and Theorem 5 below, that MYULA and PULA satisfy these conditions if at least one of the following assumptions is verified:

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

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

Note that if for any  $\theta \in \Theta$ ,  $U_\theta$  is convex on  $\mathbb{R}^d$  and  $\sup_{\theta \in \Theta}(\int_{\mathbb{R}^d} \exp[-U_\theta(\tilde{x})] d\tilde{x}) < +\infty$ , then H3 is automatically satisfied, as an immediate extension of [4, Lemma 2.2 (b)]. In [49], H3 is satisfied as soon as the prior distribution  $x \mapsto p(x|\theta)$  is log-concave and proper for any  $\theta \in \Theta$ . In [49], if the prior  $x \mapsto p(x|\theta)$  is improper for some  $\theta \in \Theta$  then we require H2 to be satisfied, i.e. for any  $y \in \mathbb{C}^{d_y}$ , there exists  $m > 0$  such that for any  $\theta \in \Theta$ ,  $x \mapsto p(x|y, \theta)$  is  $m$ -log-concave. Finally, we believe that H3 could be relaxed to the following condition: there exist  $\eta > 0$  and  $c \geq 0$  such that for any  $\theta \in \Theta$  and  $x \in \mathbb{R}^d$ ,  $\min(U_\theta(x) + V_\theta(x), \bar{U}_\theta(x) + \bar{V}_\theta(x)) \geq \eta \|x\| - c$ . In particular, this latter condition holds in the case where  $x \mapsto p(x|\theta) = \exp[-\theta^\top TV(x)]$  and  $\sup_{\theta \in \Theta}(\int_{\mathbb{R}^d} \exp[-U_\theta(\tilde{x}) + V_\theta(\tilde{x})] d\tilde{x}) < +\infty$ .

Consider for any  $m \in \mathbb{N}^*$  and  $\alpha > 0$ , the two functions  $W_m$  and  $W_\alpha$  given for any  $x \in \mathbb{R}^d$  by

$$W_m(x) = 1 + \|x\|^{2m}, \quad W_\alpha = \exp\left[\alpha\sqrt{1 + \|x\|^2}\right]. \quad (19)$$

**Theorem 4.** Assume H1 and H2 or H3. Let  $\bar{\kappa} > 1 \geq \underline{\kappa} > 1/2$ ,  $\bar{\gamma} < \min\{(2 - 1/\kappa)/L, 2/(m + L)\}$  if H2 holds and  $\bar{\gamma} < \min\{(2 - 1/\kappa)/L, \eta/(2ML)\}$  if H3 holds. Then for any  $a \in (0, 1]$ , there exist  $\bar{A}_{2,a} \geq 0$  and  $\rho_a \in (0, 1)$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $R_{\gamma, \theta}$  and  $\bar{R}_{\gamma, \theta}$  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

$$\begin{aligned} \max(\|\delta_x R_{\gamma, \theta}^n - \pi_{\gamma, \theta}\|_{W^a}, \|\delta_x \bar{R}_{\gamma, \theta}^n - \bar{\pi}_{\gamma, \theta}\|_{W^a}) &\leq \bar{A}_{2,a} \rho_a^{\gamma n} W^a(x), \\ \max(\|\delta_x R_{\gamma, \theta}^n - \delta_y R_{\gamma, \theta}^n\|_{W^a}, \|\delta_x \bar{R}_{\gamma, \theta}^n - \delta_y \bar{R}_{\gamma, \theta}^n\|_{W^a}) &\leq \bar{A}_{2,a} \rho_a^{\gamma n} \{W^a(x) + W^a(y)\}, \end{aligned}$$

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

*Proof.* The proof is postponed to Section 5.2. □

**Theorem 5.** Assume H1 and H2 or H3. Let  $\bar{\kappa} > 1 \geq \underline{\kappa} > 1/2$ ,  $\bar{\gamma} < 2/(m + L)$  if H2 holds and  $\bar{\gamma} < 2/L$  if H3 holds. Then for any  $a \in (0, 1]$ , there exist  $A_{2,a} \geq 0$  and  $\rho_a \in (0, 1)$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $S_{\gamma, \theta}$  and  $\bar{S}_{\gamma, \theta}$  admit an 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

$$\begin{aligned} \max(\|\delta_x S_{\gamma, \theta}^n - \pi_{\gamma, \theta}\|_{W^a}, \|\delta_x \bar{S}_{\gamma, \theta}^n - \bar{\pi}_{\gamma, \theta}\|_{W^a}) &\leq A_{2,a} \rho_a^{\gamma n} W^a(x), \\ \max(\|\delta_x S_{\gamma, \theta}^n - \delta_y S_{\gamma, \theta}^n\|_{W^a}, \|\delta_x \bar{S}_{\gamma, \theta}^n - \delta_y \bar{S}_{\gamma, \theta}^n\|_{W^a}) &\leq A_{2,a} \rho_a^{\gamma n} \{W^a(x) + W^a(y)\}, \end{aligned}$$

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

*Proof.* The proof is postponed to Section 5.3. □

## 4.2 Main results

We now state our main results regarding the convergence of the sequence defined by (5) under the following additional regularity assumption.

**H4.** *There exist  $M_\Theta \geq 0$  and  $f_\Theta \in C(\mathbb{R}_+, \mathbb{R}_+)$  such that for any  $\theta_1, \theta_2 \in \Theta$ ,  $x \in \mathbb{R}^d$ ,*

$$\begin{aligned} \max(\|\nabla_x V_{\theta_1}(x) - \nabla_x V_{\theta_2}(x)\|, \|\nabla_x \bar{V}_{\theta_1}(x) - \nabla_x \bar{V}_{\theta_2}(x)\|) &\leq M_\Theta \|\theta_1 - \theta_2\| (1 + \|x\|), \\ \max(\|\nabla_x U_{\theta_1}^\kappa(x) - \nabla_x U_{\theta_2}^\kappa(x)\|, \|\nabla_x \bar{U}_{\theta_1}^\kappa(x) - \nabla_x \bar{U}_{\theta_2}^\kappa(x)\|) &\leq f_\Theta(\kappa) \|\theta_1 - \theta_2\| (1 + \|x\|). \end{aligned}$$

In Theorem 6, 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 7.

**Theorem 6.** *Assume A1, A2, A3 and that  $f$  is convex. Let  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Assume H1 and one of the following conditions:*

- (a) **H2** holds,  $\bar{\gamma} < \min(2/(m+L), (2-1/\underline{\kappa})/L, L^{-1})$  and there exists  $m \in \mathbb{N}^*$  and  $C_m \geq 0$  such that for any  $\theta \in \Theta$  and  $x \in \mathbb{R}^d$ ,  $\|H_\theta(x)\| \leq C_m W_m^{1/4}(x)$  and  $\|\bar{H}_\theta(x)\| \leq C_m W_m^{1/4}(x)$ .
- (b) **H3** holds,  $\bar{\gamma} < \min((2-1/\underline{\kappa})/L, \eta/(2ML), L^{-1})$  and there exists  $0 < \alpha < \eta/4$ ,  $C_\alpha \geq 0$  such that for any  $\theta \in \Theta$  and  $x \in \mathbb{R}^d$ ,  $\|H_\theta(x)\| \leq C_\alpha W_\alpha^{1/4}(x)$  and  $\|\bar{H}_\theta(x)\| \leq C_\alpha W_\alpha^{1/4}(x)$ .

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 (5). 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 conditions 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  $\sum_{n=0}^{+\infty} \delta_{n+1}^2 \gamma_n^{-2} < +\infty$ ,  $\sum_{n=0}^{+\infty} \delta_{n+1} \gamma_{n+1}^{-3} (\gamma_n - \gamma_{n+1}) < +\infty$ .

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

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

*Proof.* The proof is postponed to Section 5.6. □

These results are similar to the ones identified in [17, 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 6, the results are stated for PULA and MYULA samplers.

Although rigorously establishing convexity of  $f$  is usually not possible for imaging models, we expect that in many cases, for any of its minimizer  $\theta_*$ ,  $f$  is convex in some neighborhood of  $\theta_*$ . For example, this is the case if its Hessian is definite positive around this point.

Assume that  $\delta_n \sim n^{-a}$ ,  $\gamma_n \sim n^{-b}$  and  $m_n \sim n^{-c}$  with  $a, b, c \geq 0$ . 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 size).

- 1) In the increasing batch size case, Theorem 6 ensures that  $(\theta_n)_{n \in \mathbb{N}}$  converges if the following inequalities are satisfied

$$a + b/2 > 1, \quad a - b + c > 1, \quad a \leq 1. \tag{20}$$

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 (20) 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 (20) hold.

- 2) In the fixed batch size case, which implies that  $c = 0$ , Theorem 6 ensures that  $(\theta_n)_{n \in \mathbb{N}}$  converges if the following inequalities are satisfied

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

which can be rewritten as

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

The interval  $(2(a-1), \min(a-1/2, a/2))$  is then not empty if and only if  $a \in (5/6, 1]$ .

**Theorem 7.** Assume **A1**, **A2**, **A3** and that  $f$  is convex. Let  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Assume **H1** and that the condition **(a)** or **(b)** in Theorem **6** 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 (5)

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

where

(a)

$$E_n = C_1 \left\{ 1 + \sum_{k=0}^{n-1} \delta_{k+1} \gamma_k^{1/2} + \sum_{k=0}^{n-1} \delta_{k+1} / (m_k \gamma_k) + \sum_{k=0}^{n-1} \delta_{k+1}^2 / (m_k \gamma_k)^2 \right\}. \quad (21)$$

(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

$$E_n = C_2 \left\{ 1 + \sum_{k=0}^{n-1} \delta_{k+1} \gamma_k^{1/2} + \sum_{k=0}^{n-1} \delta_{k+1}^2 / \gamma_k + \sum_{k=0}^{n-1} \delta_{k+1} \gamma_{k+1}^{-3} (\gamma_k - \gamma_{k+1}) \right\}. \quad (22)$$

*Proof.* The proof is postponed to Section 5.7.  $\square$

First, note that if the stepsize is fixed and recalling that  $\kappa = \lambda/\gamma$  then the condition  $\gamma < (2 - 1/\kappa)/L$  can be rewritten as  $\gamma < 2/(L + \lambda^{-1})$ . Assume that  $(\delta_n)_{n \in \mathbb{N}}$  is non-increasing,  $\lim_{n \rightarrow +\infty} \delta_n = 0$ ,  $\lim_{n \rightarrow +\infty} m_n = +\infty$  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 [37, Problem 80, Part I], it holds that

$$\begin{cases} \lim_{n \rightarrow +\infty} [(\sum_{k=1}^n \delta_k / m_k) / (\sum_{k=1}^n \delta_k)] = \lim_{n \rightarrow +\infty} 1/m_n = 0; \\ \lim_{n \rightarrow +\infty} [(\sum_{k=1}^n \delta_k^2) / (\sum_{k=1}^n \delta_k)] = \lim_{n \rightarrow +\infty} \delta_n = 0. \end{cases} \quad (23)$$

Therefore, using (21) we obtain that

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

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 (22) and (23) we obtain that

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

## 5 Proof of the main results

In this section, we gather the proofs of Section 4. First, in Section 5.1 we derive some useful technical lemmas. In Section 5.2, we prove Theorem 4, using minorisation and Foster-Lyapunov drift conditions. Similarly, we prove Theorem 5 in Section 5.3. Next, we show Theorem 6 by applying [17, Theorem 1, Theorem 3] and Theorem 7 by applying [17, Theorem 2, Theorem 4], which boils down to verifying that [17, H1, H2] are satisfied. In Section 5.4, we show that [17, H1, H2] hold if the sequence is given by (5) where  $\{(K_{\gamma,\theta}, \bar{K}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(R_{\gamma,\theta}, \bar{R}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$  defined in (18), *i.e.* we consider PULA as a sampling scheme in the optimization algorithm. In Section 5.5 we check that [17, H1, H2] are satisfied when  $\{(K_{\gamma,\theta}, \bar{K}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(S_{\gamma,\theta}, \bar{S}_{\gamma,\theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$  defined in (17), *i.e.* when considering MYULA as a sampling scheme. Finally, we prove Theorem 6 in Section 5.6 and Theorem 7 in Section 5.7.

## 5.1 Technical lemmas

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

$$RW(x) \leq \lambda W(x) + b.$$

We will use the following result.

**Lemma 8.** *Let  $R$  be a Markov kernel on  $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$  which satisfies  $\mathbf{D}_d(W, \lambda^\gamma, b\gamma)$  with  $\lambda \in (0, 1)$ ,  $b \geq 0$ ,  $\gamma > 0$  and a measurable function  $W : \mathbb{R}^d \rightarrow [1, +\infty)$ . Then, we have for any  $x \in \mathbb{R}^d$*

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

*Proof.* Using [17, Lemma 9] we have for any  $x \in \mathbb{R}^d$

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

□

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

**Lemma 9.** *Let  $\kappa > 0$  and  $U : \mathbb{R}^d \rightarrow \mathbb{R}$  convex. Assume that  $U$  is  $M$ -Lipschitz with  $M \geq 0$ , then  $U^\kappa$  is  $M$ -Lipschitz and for any  $x \in \mathbb{R}^d$ ,  $\|x - \text{prox}_U^\kappa(x)\| \leq \kappa M$ .*

*Proof.* Let  $\kappa > 0$ . We have for any  $x, y \in \mathbb{R}^d$  by (7) and (8)

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

Hence,  $U^\kappa$  is  $M$ -Lipschitz. Since by [5, Proposition 12.30],  $U^\kappa$  is continuously differentiable we have for any  $x \in \mathbb{R}^d$ ,  $\|\nabla U^\kappa(x)\| \leq M$ . Combining this result with the fact that for any  $x \in \mathbb{R}^d$ ,  $\nabla U^\kappa(x) = (x - \text{prox}_U^\kappa(x))/\kappa$  by [5, Proposition 12.30] concludes the proof. □

**Lemma 10.** *Let  $U : \mathbb{R}^d \rightarrow [0, +\infty)$  be a convex and  $M$ -Lipschitz function with  $M \geq 0$ . Then for any  $\kappa > 0$  and  $z, z' \in \mathbb{R}^d$ ,*

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

*Proof.*  $\kappa > 0$  and  $z, z' \in \mathbb{R}^d$ . Since  $(z - \text{prox}_U^\kappa(z))/\kappa \in \partial U(\text{prox}_U^\kappa(z))$  [5, Proposition 16.44], we have

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

Combining this result, the fact that  $U$  is  $M$ -Lipschitz and Lemma 9 we get that

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

which concludes the proof □

**Lemma 11.** *Let  $\kappa_1, \kappa_2 > 0$  and  $U : \mathbb{R}^d \rightarrow \mathbb{R}$  convex and lower semi-continuous. For any  $x \in \mathbb{R}^d$  we have*

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

If in addition,  $U$  is  $M$ -Lipschitz with  $M \geq 0$  then

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

*Proof.* Let  $x \in \mathbb{R}^d$ . By definition of  $\text{prox}_U^{\kappa_1}(x)$  we have

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

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

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

which concludes the proof.  $\square$

**Lemma 12.** *Let  $V : \mathbb{R}^d \rightarrow \mathbb{R}$   $\mathfrak{m}$ -convex and continuously differentiable with  $\mathfrak{m} \geqslant 0$ . Assume that there exists  $M > 0$  such that for any  $x, y \in \mathbb{R}^d$*

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

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

$$\|x - \gamma \nabla V(x)\|^2 \leqslant (1 - \gamma \varpi) \|x\|^2 + \gamma \{(2/(\mathfrak{m} + M) - \bar{\gamma})^{-1} + 4\varpi\} \|x^*\|^2,$$

*with  $\varpi = \mathfrak{m}M/(\mathfrak{m} + M)$ .*

*Proof.* Let  $x \in \mathbb{R}^d$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $\bar{\gamma} < 2/(\mathfrak{m} + M)$ . Using [36, Theorem 2.1.11] and the fact that for any  $a, b, \varepsilon > 0$ ,  $\varepsilon a^2 + b^2/\varepsilon \geqslant 2ab$  we have

$$\begin{aligned} & \|x - \gamma \nabla V(x)\|^2 \\ & \leqslant \|x\|^2 - 2\gamma \langle \nabla V(x) - \nabla V(x^*), x - x^* \rangle + \gamma \bar{\gamma} \|\nabla V(x) - \nabla V(x^*)\|^2 \\ & \quad + 2\gamma \|x^*\| \|\nabla V(x) - \nabla V(x^*)\| \\ & \leqslant \|x\|^2 - 2\gamma \varpi \|x - x^*\|^2 - \gamma(2/(\mathfrak{m} + M) - \bar{\gamma}) \|\nabla V(x) - \nabla V(x^*)\|^2 \\ & \quad + 2\gamma \|x^*\| \|\nabla V(x) - \nabla V(x^*)\| \\ & \leqslant \|x\|^2 - 2\gamma \varpi \|x - x^*\|^2 - \gamma(2/(\mathfrak{m} + M) - \bar{\gamma}) \|\nabla V(x) - \nabla V(x^*)\|^2 \\ & \quad + \gamma(2/(\mathfrak{m} + M) - \bar{\gamma}) \|\nabla V(x) - \nabla V(x^*)\|^2 + \gamma/(2/(\mathfrak{m} + M) - \bar{\gamma}) \|x^*\|^2 \\ & \leqslant (1 - 2\gamma \varpi) \|x\|^2 + 4\gamma \varpi \|x^*\| \|x\| + \gamma/(2/(\mathfrak{m} + M) - \bar{\gamma}) \|x^*\|^2 \\ & \leqslant (1 - \gamma \varpi) \|x\|^2 + \gamma \{(2/(\mathfrak{m} + M) - \bar{\gamma})^{-1} + 4\varpi\} \|x^*\|^2. \end{aligned}$$

$\square$

**Lemma 13.** *Assume **H1** and **H2**. Then for any  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\gamma} < 2/(\mathfrak{m} + L)$  and  $x \in \mathbb{R}^d$ , we have*

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

*with  $\varpi = \mathfrak{m}L/(\mathfrak{m} + L)$ .*

*Proof.* Let  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Using **H1**, **H2**, Lemma 9, Lemma 12, 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

$$\begin{aligned} & \|\text{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\text{prox}_{U_\theta}^{\gamma\kappa}(x))\|^2 \\ & \leqslant (1 - \gamma \varpi) \|\text{prox}_{U_\theta}^{\gamma\kappa}(x)\|^2 + \gamma \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} \|x_\theta^*\|^2 \\ & \leqslant (1 - \gamma \varpi) \|x - \text{prox}_{U_\theta}^{\gamma\kappa}(x)\|^2 + \gamma \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 \\ & \leqslant (1 - \gamma \varpi) \|x\|^2 + \gamma^2 \kappa^2 M^2 + 2\gamma \kappa M \|x\| + \gamma \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 \\ & \leqslant (1 - \gamma \varpi/2) \|x\|^2 + \gamma^2 \kappa^2 M^2 + \gamma \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 + 2\gamma \kappa M \|x\| - \gamma \varpi \|x\|^2 / 2 \\ & \leqslant (1 - \gamma \varpi/2) \|x\|^2 + \gamma \bar{\gamma} \kappa^2 M^2 + \gamma \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 + 2\gamma \kappa^2 M^2 \varpi^{-1}. \end{aligned}$$

$\square$

**Lemma 14.** Assume **H1** and **H3**. Then for any  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\gamma} < 2/\mathbf{L}$  and  $x \in \mathbb{R}^d$ , we have

$$\begin{aligned} \|\text{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\text{prox}_{U_\theta}^{\gamma\kappa}(x))\|^2 &\leq \|x\|^2 + \gamma [3\bar{\gamma}\kappa^2\mathbf{M}^2 + 2\kappa\mathbf{c} + 2\kappa(R_{U,2} + \mathbf{M}R_{U,1}) \\ &\quad + (2/L - \bar{\gamma})^{-1}R_{V,1}^2 - 2\kappa\eta\|x\|] . \end{aligned}$$

*Proof.* Let  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Using **H1**, **H3**, Lemma 9 and Lemma 10 and Lemma 12 we have

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

□

**Lemma 15.** Assume **H1** and **H2**. Then for any  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\gamma} < 2/(\mathbf{m} + \mathbf{L})$  and  $x \in \mathbb{R}^d$ , we have

$$\begin{aligned} \|x - \gamma \nabla_x V_\theta(x) - \gamma \nabla_x U_\theta^{\gamma\kappa}(x)\|^2 &\leq (1 - \gamma\varpi/2)\|x\|^2 \\ &\quad + \gamma \{(2/(\mathbf{m} + \mathbf{L}) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 + 2\gamma^2\mathbf{M}R_{V,1} + \gamma^2\mathbf{M}^2 + 2\gamma\mathbf{M}^2(1 + \bar{\gamma}\mathbf{L})^2\varpi^{-1} , \end{aligned}$$

with  $\varpi = \mathbf{m}\mathbf{L}/(2\mathbf{m} + 2\mathbf{L})$ .

*Proof.* Let  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Using **H1**, **H2**, Lemma 9, Lemma 12 and that for any  $\alpha, \beta \geq 0$ ,  $\max(-\alpha t^2 + 2\beta t) = \beta^2/\alpha$  we have

$$\begin{aligned} \|x - \gamma \nabla_x V_\theta(x) - \gamma \nabla_x U_\theta^{\gamma\kappa}(x)\|^2 &\leq \|x - \gamma \nabla_x V_\theta(x)\|^2 + 2\gamma\mathbf{M}\|x - \gamma\{\nabla_x V_\theta(x) - \nabla_x V_\theta(x_\theta^*)\}\| + \gamma^2\mathbf{M}^2 \\ &\leq (1 - \gamma\varpi)\|x\|^2 + \gamma \{(2/(\mathbf{m} + \mathbf{L}) - \bar{\gamma})^{-1} + 4\varpi\} \|x_\theta^*\|^2 \\ &\quad + 2\gamma\mathbf{M}\|x\| + 2\gamma^2\mathbf{M}\|\nabla_x V_\theta(x) - \nabla_x V_\theta(x_\theta^*)\| + \gamma^2\mathbf{M}^2 \\ &\leq (1 - \gamma\varpi)\|x\|^2 + \gamma \{(2/(\mathbf{m} + \mathbf{L}) - \bar{\gamma})^{-1} + 4\varpi\} \|x_\theta^*\|^2 \\ &\quad + 2\gamma\mathbf{M}\|x\| + 2\gamma^2\mathbf{M}\|x\| + 2\gamma^2\mathbf{M}\|x_\theta^*\| + \gamma^2\mathbf{M}^2 \\ &\leq (1 - \gamma\varpi/2)\|x\|^2 + \gamma \{(2/(\mathbf{m} + \mathbf{L}) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 \\ &\quad + 2\gamma^2\mathbf{M}R_{V,1} + \gamma^2\mathbf{M}^2 + 2\gamma\mathbf{M}(1 + \bar{\gamma}\mathbf{L})\|x\| - \gamma\varpi\|x\|^2/2 \\ &\leq (1 - \gamma\varpi/2)\|x\|^2 + \gamma \{(2/(\mathbf{m} + \mathbf{L}) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 \\ &\quad + 2\gamma^2\mathbf{M}R_{V,1} + \gamma^2\mathbf{M}^2 + 2\gamma\mathbf{M}^2(1 + \bar{\gamma}\mathbf{L})^2\varpi^{-1} . \end{aligned}$$

□

**Lemma 16.** Assume **H1** and **H3**. Then for any  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $x \in \mathbb{R}^d$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\gamma} < \min(2/\mathbf{L}, \eta/(2\mathbf{M}\mathbf{L}))$ , we have

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

*Proof.* Let  $\kappa > 0$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Using **H1**, **H3**, (7), Lemma 9 and Lemma 10 we

have

$$\begin{aligned}
& \|x - \gamma \nabla_x V_\theta(x) - \gamma \nabla_x U_\theta^{\gamma\kappa}(x)\|^2 \\
& \leq \|x - \gamma \nabla_x V_\theta(x)\|^2 - 2\gamma \langle x - \gamma \nabla_x V_\theta(x), \nabla_x U_\theta^{\gamma\kappa}(x) \rangle + \gamma^2 M^2 \\
& \leq \|x - \gamma \nabla_x V_\theta(x)\|^2 - 2\kappa^{-1} \langle x - \gamma \nabla_x V_\theta(x), x - \text{prox}_{U_\theta}^{\gamma\kappa}(x) \rangle + \gamma^2 M^2 \\
& \leq \|x - \gamma \nabla_x V_\theta(x)\|^2 - 2\kappa^{-1} \langle x, x - \text{prox}_{U_\theta}^{\gamma\kappa}(x) \rangle + 2\kappa^{-1} \gamma \|\nabla_x V_\theta(x)\| \|x - \text{prox}_{U_\theta}^{\gamma\kappa}(x)\| + \gamma^2 M^2 \\
& \leq \|x - \gamma \nabla_x V_\theta(x)\|^2 + 3\gamma^2 M^2 - 2\gamma\eta \|x\| + 2\gamma c + 2\gamma(M\|x_\theta^\sharp\| + U(x_\theta^\sharp)) + 2\gamma\bar{\gamma}M \|\nabla_x V_\theta(x)\| \\
& \leq \|x - \gamma \nabla_x V_\theta(x)\|^2 + 3\gamma\bar{\gamma}M^2 - 2\gamma\eta \|x\| \\
& \quad + 2\gamma c + 2\gamma(MR_{U,1} + R_{U,2}) + 2\gamma\bar{\gamma}ML \|x\| + 2\gamma\bar{\gamma}ML \|x_\theta^*\| \\
& \leq \|x - \gamma \nabla_x V_\theta(x)\|^2 + 3\gamma\bar{\gamma}M^2 - \gamma\eta \|x\| + 2\gamma c + 2\gamma(MR_{U,1} + R_{U,2}) + 2\gamma\bar{\gamma}ML \|x_\theta^*\| ,
\end{aligned}$$

where we have used for the last inequality that  $\bar{\gamma} < \eta/(2ML)$ . Then, we can conclude using **H1** and Lemma 12 that

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

□

For  $v \in \mathbb{R}^d$  and  $\sigma > 0$ , denote  $\Upsilon_{v,\sigma}$  the  $d$ -dimensional Gaussian distribution with mean  $v$  and covariance matrix  $\sigma^2 \text{Id}$ .

**Lemma 17.** *For any  $\sigma_1, \sigma_2 > 0$  and  $v_1, v_2 \in \mathbb{R}^d$ , we have*

$$\text{KL}(\Upsilon_{v_1, \sigma_1} \text{Id} | \Upsilon_{v_2, \sigma_2} \text{Id}) = \|v_1 - v_2\|^2 / (2\sigma_2^2) + (d/2) \{-\log(\sigma_1^2/\sigma_2^2) - 1 + \sigma_1^2/\sigma_2^2\} .$$

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

$$\text{KL}(\Upsilon_{v_1, \sigma_1} \text{Id} | \Upsilon_{v_2, \sigma_2} \text{Id}) \leq \|v_1 - v_2\|^2 / (2\sigma_2^2) + (d/2)(1 - \sigma_1^2/\sigma_2^2)^2 .$$

*Proof.* Let  $X$  be a  $d$ -dimensional Gaussian random variable with mean  $v_1$  and covariance matrix  $\sigma_1^2 \text{Id}$ . We have that

$$\begin{aligned}
\text{KL}(\Upsilon_{v_1, \sigma_1} \text{Id} | \Upsilon_{v_2, \sigma_2} \text{Id}) &= \mathbb{E} \left[ \log \left\{ (\sigma_2^2/\sigma_1^2)^{d/2} \exp \left[ -\|X - v_1\|^2 / (2\sigma_1^2) + \|X - v_2\|^2 / (2\sigma_2^2) \right] \right\} \right] \\
&= -(d/2) \log(\sigma_1^2/\sigma_2^2) + \mathbb{E} \left[ -\|X - v_1\|^2 / (2\sigma_1^2) + \|X - v_2\|^2 / (2\sigma_2^2) \right] \\
&= -(d/2) \log(\sigma_1^2/\sigma_2^2) + (1/2)(\sigma_2^{-2} - \sigma_1^{-2}) \mathbb{E} \left[ -\|X - v_1\|^2 \right] + \|v_1^2 - v_2^2\| / (2\sigma_2^2) \\
&= -(d/2) \log(\sigma_1^2/\sigma_2^2) + (d/2)(\sigma_1^2/\sigma_2^2 - 1) + \|v_1^2 - v_2^2\| / (2\sigma_2^2) \\
&= \|v_1 - v_2\|^2 / (2\sigma_2^2) + (d/2) \{-\log(\sigma_1^2/\sigma_2^2) - 1 + \sigma_1^2/\sigma_2^2\} .
\end{aligned}$$

In the case where  $\sigma_1 \geq \sigma_2$ , let  $s = \sigma_1^2/\sigma_2^2 - 1$ . Since  $s \geq 0$  we have  $\log(1+s) \geq s - s^2$ . Therefore, we get that

$$-\log(\sigma_1^2/\sigma_2^2) - 1 + \sigma_1^2/\sigma_2^2 = -\log(1+s) + s \leq s^2 ,$$

which concludes the proof. □

## 5.2 Proof of Theorem 4

We show that under **H2** or **H3**, Foster-Lyapunov drifts hold for MYULA in Lemma 18 and Lemma 19. Combining these Foster-Lyapunov drifts with an appropriate minorisation condition Lemma 20, we obtain the geometric ergodicity of the underlying Markov chain in Theorem 21.

**Lemma 18.** *Assume **H1** and **H2**. Then for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \kappa > 1/2$ ,  $\bar{\gamma} < 2/(\mathfrak{m} + L)$ ,  $R_{\gamma, \theta}$  and  $\bar{R}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W_1, \lambda_2^\gamma, b_2\gamma)$  with*

$$\lambda_2 = \exp[-\varpi/2] ,$$

$$b_2 = \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 + 2\bar{\gamma}MLR_{V,1} + \bar{\gamma}M^2 + 2d + 2M^2(1 + \bar{\gamma}L)^2\varpi^{-1} + \varpi/2 ,$$

$$\varpi = \mathfrak{m}L/(\mathfrak{m} + L) ,$$

where for any  $x \in \mathbb{R}^d$ ,  $W_2(x) = 1 + \|x\|^2$ . In addition, for any  $m \in \mathbb{N}^*$ , there exist  $\lambda_m \in (0, 1)$ ,  $b_m \geq 0$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ ,  $\bar{\gamma} < 2/(\mathbf{m} + \mathbf{L})$ ,  $\mathbf{R}_{\gamma, \theta}$  and  $\bar{\mathbf{R}}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W_m, \lambda_m^\gamma, b_m \gamma)$ , where  $W_m$  is given in (19).

*Proof.* We show the property for  $\mathbf{R}_{\gamma, \theta}$  only as the proof for  $\bar{\mathbf{R}}_{\gamma, \theta}$  is identical. Let  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Let  $Z$  be a  $d$ -dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 15 we have

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

Therefore, we get

$$\begin{aligned} \int_{\mathbb{R}^d} (1 + \|y\|^2) \mathbf{R}_{\gamma, \theta}(x, dy) &\leq (1 - \gamma \varpi/2)(1 + \|x\|^2) + \gamma \left[ \left\{ (2/(\mathbf{m} + \mathbf{L}) - \bar{\gamma})^{-1} + 4\varpi \right\} R_{V,1}^2 \right. \\ &\quad \left. + 2\bar{\gamma} \mathbf{M} L R_{V,1} + \bar{\gamma} \mathbf{M}^2 + 2d + 2\mathbf{M}^2(1 + \bar{\gamma} \mathbf{L})^2 \varpi^{-1} + \varpi/2 \right], \end{aligned}$$

which concludes the first part of the proof. Let  $\mathcal{T}_{\gamma, \theta}(x) = x - \gamma \nabla_x V_\theta(x) - \gamma \nabla_x U_\theta^{\gamma \kappa}(x)$ . In the sequel, for any  $k \in \{1, \dots, m\}$ ,  $b, \tilde{b}_k \geq 0$  and  $\lambda, \tilde{\lambda}_k \in [0, 1]$  are constants independent of  $\gamma$  which may take different values at each appearance. Note that using Lemma 15, for any  $k \in \{1, \dots, 2m\}$  there exist  $\tilde{\lambda}_k \in (0, 1)$  and  $\tilde{b}_k \geq 0$  such that

$$\begin{aligned} \|\mathcal{T}_{\gamma, \theta}(x)\|^k &\leq \{\tilde{\lambda}_k^\gamma \|x\| + \gamma \tilde{b}_k\}^k \\ &\leq \tilde{\lambda}_k^\gamma \|x\|^k + \gamma 2^k \max(\tilde{b}_k, 1)^k \max(\bar{\gamma}, 1)^{2k-1} \left\{ 1 + \|x\|^{k-1} \right\} \\ &\leq \tilde{\lambda}_k^\gamma \|x\|^k + \tilde{b}_k \gamma \left\{ 1 + \|x\|^{k-1} \right\} \leq (1 + \|x\|^k)(1 + \tilde{b}_k \gamma). \end{aligned} \tag{24}$$

Therefore, combining (24) and the Cauchy-Schwarz inequality we obtain

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

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

**Lemma 19.** Assume **H1** and **H3**. Then for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ ,

$\underline{\kappa} > 1/2$ ,  $\bar{\gamma} < \min(2/\mathbf{L}, \eta/(2\mathbf{M}\mathbf{L}))$ ,  $R_{\gamma,\theta}$  and  $\bar{R}_{\gamma,\theta}$  satisfy  $\mathbf{D}_d(W, \lambda^\gamma, b\gamma)$  with

$$\begin{aligned} \lambda &= e^{-\alpha^2}, \\ b_e &= (4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + (3/2)\bar{\gamma}\mathbf{M}^2 + \mathbf{c} + \mathbf{M}R_{U,1} + R_{U,2} + \bar{\gamma}\mathbf{M}\mathbf{L}R_{V,2} + d + 2\alpha, \\ b &= \alpha b_e e^{\alpha\bar{\gamma}b_e} W(R), \\ W &= W_\alpha, \quad \alpha < \eta/8, \\ R_\eta &= \max(2b_e/(\eta - 8\alpha), 1), \end{aligned} \tag{25}$$

where  $W_\alpha$  is given in (19).

*Proof.* We show the property for  $R_{\gamma,\theta}$  only as the proof for  $\bar{R}_{\gamma,\theta}$  is identical. Let  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $x \in \mathbb{R}^d$  and  $Z$  be a  $d$ -dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 16 we have

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

Using the log-Sobolev inequality [3, Proposition 5.4.1] and Jensen's inequality we get that

$$\begin{aligned} R_{\gamma,\theta}W(x) &\leq \exp[\alpha R_{\gamma,\theta}\phi(x) + \alpha^2\gamma] \\ &\leq \exp\left[\alpha\left(1 + \int_{\mathbb{R}^d} \|y\|^2 R_{\gamma,\theta}(x, dy)\right)^{1/2} + \alpha^2\gamma\right]. \end{aligned} \tag{26}$$

We now distinguish two cases:

(a) If  $\|x\| \geq R_\eta$ , recalling that  $R_\eta$  is given in (25), then

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

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

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

Hence,

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

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

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

Therefore, using (26), we get

$$\begin{aligned} R_{\gamma,\theta}W(x) &\leq \\ &\leq \exp[\alpha\gamma \{(4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + (3/2)\bar{\gamma}\mathbf{M}^2 + \mathbf{c} + \mathbf{M}R_{U,1} + R_{U,2} + \bar{\gamma}\mathbf{M}\mathbf{L}R_{V,2} + d + \alpha\}] W(x). \end{aligned}$$

Since for all  $a \geq b$ ,  $e^a - e^b \leq (a-b)e^a$  we obtain that

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

which concludes the proof.  $\square$

**Lemma 20.** Assume **H1**. For any  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ ,  $\bar{\gamma} < (2 - 1/\underline{\kappa})/\mathbf{L}$  and  $x, y \in \mathbb{R}^d$

$$\max \left( \|\delta_x R_{\gamma, \theta}^{[1/\gamma]} - \delta_y R_{\gamma, \theta}^{[1/\gamma]}\|_{\text{TV}}, \|\delta_x \bar{R}_{\gamma, \theta}^{[1/\gamma]} - \delta_y \bar{R}_{\gamma, \theta}^{[1/\gamma]}\|_{\text{TV}} \right) \leq 1 - 2\Phi \left\{ - \|x - y\| / (2\sqrt{2}) \right\},$$

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

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

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

Combining this result with [36, Theorem 2.1.5, Equation (2.1.8)] and the fact that  $\gamma \leq 2/\{\mathbf{L} + 1/(\gamma\kappa)\}$  since  $\bar{\gamma} \leq (2 - 1/\underline{\kappa})/\mathbf{L}$ , we have for any  $x, y \in \mathbb{R}^d$

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

The proof is then an application of [16, Proposition 3b] with  $\ell \leftarrow 1$ , for any  $x \in \mathbb{R}^d$ ,  $\mathcal{T}_{\gamma, \theta}(x) \leftarrow x - \gamma \nabla_x V_\theta(x) - \gamma \nabla_x U_\theta^{\gamma\kappa}(x)$  and  $\Pi \leftarrow \text{Id}$ .  $\square$

**Theorem 21.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ ,  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/\mathbf{L}, 2/(\mathbf{m} + \mathbf{L})\}$  if **H2** holds and  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/\mathbf{L}, \eta/(2\mathbf{M}\mathbf{L})\}$  if **H3** holds. Then for any  $a \in (0, 1]$ , there exist  $A_{2,a} \geq 0$  and  $\rho_a \in (0, 1)$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $R_{\gamma, \theta}$  and  $\bar{R}_{\gamma, \theta}$  admit invariant probability measures  $\pi_{\gamma, \theta}$ , respectively  $\bar{\pi}_{\gamma, \theta}$ , and for any  $x, y \in \mathbb{R}^d$  and  $n \in \mathbb{N}$  we have

$$\begin{aligned} \max \left( \|\delta_x R_{\gamma, \theta}^n - \pi_{\gamma, \theta}\|_{W^a}, \|\delta_x \bar{R}_{\gamma, \theta}^n - \bar{\pi}_{\gamma, \theta}\|_{W^a} \right) &\leq A_{2,a} \rho_a^{\gamma n} W^a(x), \\ \max \left( \|\delta_x R_{\gamma, \theta}^n - \delta_y R_{\gamma, \theta}^n\|_{W^a}, \|\delta_x \bar{R}_{\gamma, \theta}^n - \delta_y \bar{R}_{\gamma, \theta}^n\|_{W^a} \right) &\leq A_{2,a} \rho_a^{\gamma n} \{W^a(x) + W^a(y)\}, \end{aligned}$$

with  $W = W_m$  and  $m \in \mathbb{N}^*$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta/8)$  if **H3** holds, see (19).

*Proof.* We only show that for any  $a \in (0, 1]$ , there exist  $A_{2,a} \geq 0$  and  $\rho_a \in (0, 1)$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  we have  $\|\delta_x R_{\gamma, \theta}^n - \pi_{\gamma, \theta}\|_{W^a} \leq A_{2,a} \rho_a^{\gamma n} W^a(x)$  and  $\|\delta_x R_{\gamma, \theta}^n - \delta_y R_{\gamma, \theta}^n\|_{W^a} \leq A_{2,a} \rho_a^{\gamma n} \{W^a(x) + W^a(y)\}$ , since the proof for  $\bar{R}_{\gamma, \theta}$  is similar. Let  $a \in [0, 1]$ . First, using Jensen's inequality and Lemma 18 if **H2** holds or Lemma 19 if **H3** holds, we get that there exist  $\lambda_a$  and  $b_a$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $R_{\gamma, \theta}$  and  $\bar{R}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W^a, \lambda_a^\gamma, b_a \gamma)$ . Combining [16, Theorem 6], Lemma 20 and  $\mathbf{D}_d(W^a, \lambda_a^\gamma, b_a \gamma)$ , we get that there exist  $\bar{A}_{2,a} \geq 0$  and  $\rho_a \in (0, 1)$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $x, y \in \mathbb{R}^d$  and  $n \in \mathbb{N}$ ,  $R_{\gamma, \theta}$  and  $\bar{R}_{\gamma, \theta}$  admit invariant probability measures  $\pi_{\gamma, \theta}$  and  $\bar{\pi}_{\gamma, \theta}$  respectively and

$$\max \left\{ \|\delta_x R_{\gamma, \theta}^n - \delta_y R_{\gamma, \theta}^n\|_{W^a}, \|\delta_x \bar{R}_{\gamma, \theta}^n - \delta_y \bar{R}_{\gamma, \theta}^n\|_{W^a} \right\} \leq \bar{A}_{2,a} \rho_a^{\gamma n} \{W^a(x) + W^a(y)\}. \quad (27)$$

Using that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$ ,  $R_{\gamma, \theta}$  and  $\bar{R}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W^a, \lambda_a^\gamma, b_a \gamma)$  and [17, Lemma S2] we have

$$\pi_{\gamma, \theta}(W^a) \leq b_a \gamma / (1 - \lambda_a^\gamma) \leq b_a \lambda_a^{-\bar{\gamma}} / \log(1/\lambda_a). \quad (28)$$

Hence, combining (27) and (28), we have for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $n \in \mathbb{N}$

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

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

### 5.3 Proof of Theorem 5

We show that under **H2** or **H3**, Foster-Lyapunov drifts hold for PULA in Lemma 22 and Lemma 23. Combining these Foster-Lyapunov drifts with an appropriate minorisation condition Lemma 24, we obtain the geometric ergodicity of the underlying Markov chain in Theorem 25.

**Lemma 22.** Assume **H1** and **H2**. Then for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$  and  $\bar{\gamma} < 2/(\mathfrak{m} + L)$ ,  $S_{\gamma, \theta}$  and  $\bar{S}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W_1, \lambda_2^\gamma, b_2\gamma)$  with

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

where for any  $x \in \mathbb{R}^d$ ,  $W_1(x) = 1 + \|x\|^2$ . In addition, for any  $m \in \mathbb{N}^*$ , there exist  $\lambda_m \in (0, 1)$ ,  $b_m \geq 0$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$  and  $\bar{\gamma} < 2/(\mathfrak{m} + L)$ ,  $S_{\gamma, \theta}$  and  $\bar{S}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W_m, \lambda_m^\gamma, b_m\gamma)$ , where  $W_m$  is given in (19).

*Proof.* We show the property for  $S_{\gamma, \theta}$  only as the proof for  $\bar{S}_{\gamma, \theta}$  is identical. Let  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Let  $Z$  be a  $d$ -dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 13 we have

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

Therefore, we get

$$\begin{aligned}\int_{\mathbb{R}^d} (1 + \|y\|^2) S_{\gamma, \theta}(x, dy) &\leq (1 - \gamma\varpi/2)(1 + \|x\|^2) + \gamma [\bar{\gamma}\bar{\kappa}^2M^2 \\ &\quad + \{(2/(\mathfrak{m} + L) - \bar{\gamma})^{-1} + 4\varpi\} R_{V,1}^2 + 2d + 2\bar{\kappa}^2M^2\varpi^{-1} + \varpi/2] ,\end{aligned}$$

which concludes the first part of the proof using that for any  $t \geq 0$ ,  $1 - t \leq e^{-t}$ . The proof of the result for  $W = W_m$  with  $m \in \mathbb{N}^*$  is a straightforward adaptation of the one of Lemma 18 and is left to the reader.  $\square$

**Lemma 23.** Assume **H1** and **H3**. Then for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$  and  $\bar{\gamma} < 2/L$ ,  $S_{\gamma, \theta}$  and  $\bar{S}_{\gamma, \theta}$  satisfy  $\mathbf{D}_d(W, \lambda^\gamma, b\gamma)$  with

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

and where  $W_\alpha$  is given in (19).

*Proof.* We show the property for  $S_{\gamma, \theta}$  only as the proof for  $\bar{S}_{\gamma, \theta}$  is identical. Let  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $x \in \mathbb{R}^d$ , and  $Z$  be a  $d$ -dimensional Gaussian random variable with zero mean and identity covariance matrix. Using Lemma 14 we have

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

Using the log-Sobolev inequality [3, Proposition 5.4.1] and Jensen's inequality we get that

$$\begin{aligned}S_{\gamma, \theta}W(x) &\leq \exp[\alpha S_{\gamma, \theta}\phi(x) + \alpha^2\gamma] \\ &\leq \exp \left[ \alpha \left( 1 + \int_{\mathbb{R}^d} \|y\|^2 S_{\gamma, \theta}(x, dy) \right)^{1/2} + \alpha^2\gamma \right] .\end{aligned}\tag{29}$$

We now distinguish two cases.

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

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

Hence,

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

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

$$\begin{aligned} & \left( 1 + \int_{\mathbb{R}^d} \|y\|^2 S_{\gamma,\theta}(x, dy) \right)^{1/2} - \phi(x) \\ & \leq \gamma [(3/2)\bar{\gamma}\kappa^2 M^2 + \kappa c + \kappa(R_{U,2} + M R_{U,1}) + (4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + d]. \end{aligned}$$

Therefore we get using (29)

$$\begin{aligned} & S_{\gamma,\theta}W(x)/W(x) \\ & \leq \exp [\alpha\gamma \{(3/2)\bar{\gamma}\kappa^2 M^2 + \kappa c + \kappa(R_{U,2} + M R_{U,1}) + (4/L - 2\bar{\gamma})^{-1} R_{V,1}^2 + d + \alpha\}] \leq e^{\alpha b_e \gamma}. \end{aligned}$$

Since for all  $a \geq b$ ,  $e^a - e^b \leq (a - b)e^a$  we obtain that

$$S_{\gamma,\theta}W(x) \leq \lambda^\gamma W(x) + \gamma \alpha b_e e^{\alpha \bar{\gamma} b_e} W(R_\eta),$$

which concludes the proof.  $\square$

**Lemma 24.** Assume **H1**. For any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ ,  $\bar{\gamma} < 2/L$  and  $x, y \in \mathbb{R}^d$

$$\max \left( \|\delta_x S_{\gamma,\theta}^{[1/\gamma]} - \delta_y S_{\gamma,\theta}^{[1/\gamma]}\|_{\text{TV}}, \|\delta_x \bar{S}_{\gamma,\theta}^{[1/\gamma]} - \delta_y \bar{S}_{\gamma,\theta}^{[1/\gamma]}\|_{\text{TV}} \right) \leq 1 - 2\Phi \left\{ -\|x - y\|/(2\sqrt{2}) \right\},$$

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

*Proof.* We only show that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$  with  $\bar{\gamma} < 2/L$ , and  $x, y \in \mathbb{R}^d$ ,  $\|\delta_x S_{\gamma,\theta}^{[1/\gamma]} - \delta_y S_{\gamma,\theta}^{[1/\gamma]}\|_{\text{TV}} \leq 1 - 2\Phi \{-\|x - y\|/(2\sqrt{2})\}$  since the proof for  $\bar{S}_{\gamma,\theta}$  is similar. Let  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ . Using [36, Theorem 2.1.5, Equation (2.1.8)] and that the proximal operator is non-expansive [5, Proposition 12.28], we have for any  $x, y \in \mathbb{R}^d$

$$\begin{aligned} & \|\text{prox}_{U_\theta}^{\gamma\kappa}(x) - \text{prox}_{U_\theta}^{\gamma\kappa}(y) - \gamma(\nabla_x V_\theta(\text{prox}_{U_\theta}^{\gamma\kappa}(x)) - \nabla_x V_\theta(\text{prox}_{U_\theta}^{\gamma\kappa}(y)))\| \\ & \leq \|\text{prox}_{U_\theta}^{\gamma\kappa}(x) - \text{prox}_{U_\theta}^{\gamma\kappa}(y)\| \leq \|x - y\|. \end{aligned}$$

The proof is then an application of [16, Proposition 3b] with  $\ell \leftarrow 1$ , for any  $x \in \mathbb{R}^d$ ,  $\mathcal{T}_{\gamma,\theta}(x) \leftarrow \text{prox}_{U_\theta}^{\gamma\kappa}(x) - \gamma \nabla_x V_\theta(\text{prox}_{U_\theta}^{\gamma\kappa}(x))$  and  $\Pi \leftarrow \text{Id}$ .  $\square$

**Theorem 25.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < 2/(m+L)$  if **H2** holds and  $\bar{\gamma} < 2/L$  if **H3** holds. Then for any  $a \in (0, 1]$ , there exist  $A_{2,a} \geq 0$  and  $\rho_a \in (0, 1)$  such that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $S_{\gamma,\theta}$  and  $\bar{S}_{\gamma,\theta}$  admit an invariant probability measure  $\pi_{\gamma,\theta}$  and  $\bar{\pi}_{\gamma,\theta}$  respectively, and for any  $x, y \in \mathbb{R}^d$  and  $n \in \mathbb{N}$  we have

$$\begin{aligned} & \max (\|\delta_x S_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_{W^a}, \|\delta_x \bar{S}_{\gamma,\theta}^n - \bar{\pi}_{\gamma,\theta}\|_{W^a}) \leq A_{2,a} \rho_a^{\gamma n} W^a(x), \\ & \max (\|\delta_x S_{\gamma,\theta}^n - \delta_y S_{\gamma,\theta}^n\|_{W^a}, \|\delta_x \bar{S}_{\gamma,\theta}^n - \delta_y \bar{S}_{\gamma,\theta}^n\|_{W^a}) \leq A_{2,a} \rho_a^{\gamma n} \{W^a(x) + W^a(y)\}, \end{aligned}$$

with  $W = W_m$  and  $m \in \mathbb{N}^*$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \underline{\kappa}\eta/4$  if **H3** holds, see (19).

*Proof.* The proof is similar to the one of Theorem 21.  $\square$

#### 5.4 Checking [17, H1, H2] for PULA

Lemma 26 implies that [17, H1a] holds. The geometric ergodicity proved in Theorem 25 implies [17, H1b]. Then, we show that the distance between the invariant probability distribution of the Markov chain and the target distribution is controlled in Corollary 31 and therefore [17, H1c] is satisfied. Finally, we show that [17, H2] is satisfied in Proposition 32.

**Lemma 26.** Assume **H1**, **H2** or **H3**, and let  $(X_k^n, \bar{X}_k^n)_{n \in \mathbb{N}, k \in \{0, \dots, m_n\}}$  be given by (5) with  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(S_{\gamma, \theta}, \bar{S}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$  and  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Then there exists  $A_1 \geq 1$  such that for any  $n, p \in \mathbb{N}$  and  $k \in \{0, \dots, m_n\}$

$$\begin{aligned}\mathbb{E} [S_{\gamma_n, \theta_n}^p W(X_k^n) | X_0^0] &\leq A_1 W(X_0^0), \\ \mathbb{E} [\bar{S}_{\gamma_n, \theta_n}^p W(\bar{X}_k^n) | \bar{X}_0^0] &\leq A_1 W(\bar{X}_0^0), \\ \mathbb{E} [W(X_0^0)] &< +\infty, \quad \mathbb{E} [W(\bar{X}_0^0)] < +\infty,\end{aligned}$$

with  $W = W_m$  with  $m \in \mathbb{N}^*$  and  $\bar{\gamma} < 2/(\mathfrak{m} + L)$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \underline{\kappa}\eta/4$  and  $\bar{\gamma} < 2/L$  if **H3** holds, see (19).

*Proof.* Combining [17, Lemma S15] and Lemma 22 if **H2** holds or Lemma 23 if **H3** holds conclude the proof.  $\square$

**Lemma 27.** 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, see (19).

*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

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

Hence using **H1-(a)** we have

$$\begin{aligned}\sup_{\theta \in \Theta} \pi_\theta(W) &\leq \exp [R_{V,3} + \mathfrak{m} R_{V,1}^2 / 2] \int_{\mathbb{R}^d} (1 + \|x\|^{2m}) \exp \left[ \mathfrak{m} R_{V,1} \|x\| - \mathfrak{m} \|x\|^2 / 2 \right] dx \\ &\quad \left/ \inf_{\theta \in \Theta} \left\{ \int_{\mathbb{R}^d} \exp [-U_\theta(x) - V_\theta(x)] dx \right\} \right. < +\infty.\end{aligned}$$

(b) if **H3** holds then we have

$$\begin{aligned}\int_{\mathbb{R}^d} \exp [\alpha \phi(x)] \exp [-U_\theta(x) - V_\theta(x)] dx &\leq \int_{\mathbb{R}^d} \exp [\alpha \phi(x)] \exp [-U_\theta(x)] dx \\ &\leq e^\alpha \int_{\mathbb{R}^d} \exp [\alpha(1 + \|x\|)] \exp [-\eta \|x\|] dx.\end{aligned}$$

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

$$\begin{aligned}\sup_{\theta \in \Theta} \pi_\theta(W) &\leq e^\alpha \int_{\mathbb{R}^d} \exp [\alpha(1 + \|x\|)] \exp [-\eta \|x\|] dx \\ &\quad \left/ \inf_{\theta \in \Theta} \left\{ \int_{\mathbb{R}^d} \exp [-U_\theta(x) - V_\theta(x)] dx \right\} \right. < +\infty,\end{aligned}$$

which concludes the proof.  $\square$

**Theorem 28.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < 2/(\mathfrak{m} + L)$  if **H2** holds and  $\bar{\gamma} < 2/L$  if **H3** holds. Then for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$  we have

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

where for any  $\theta \in \Theta$  and  $\gamma \in (0, \bar{\gamma}]$ ,  $\pi_{\gamma, \theta}^\#$ , respectively  $\bar{\pi}_{\gamma, \theta}^\#$ , is the invariant probability measure of  $S_{\gamma, \theta}$ , respectively  $\bar{S}_{\gamma, \theta}$ , given by (18) and associated with  $\kappa = 1$ . In addition, for any  $\gamma \in (0, \bar{\gamma}]$

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

and where  $W = W_m$  with  $m \in \mathbb{N}^*$  and  $\bar{\gamma}, \lambda, b$  are given in Lemma 22 if **H2** holds and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta)$  and  $\bar{\gamma}, \lambda, b$  are given in Lemma 23 if **H3** holds, see (19).

*Proof.* We only show that for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$ ,  $\|\pi_{\gamma, \theta}^\sharp - \pi_\theta\|_{W^{1/2}} \leq \tilde{\Psi}(\gamma)$ , since the proof of  $\|\tilde{\pi}_{\gamma, \theta}^\sharp - \tilde{\pi}_\theta\|_{W^{1/2}} \leq \tilde{\Psi}(\gamma)$  is similar. Let  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $x \in \mathbb{R}^d$ . Using Theorem 25 we obtain that  $(\delta_x S_{\gamma, \theta}^n)_{n \in \mathbb{N}}$ , with  $\kappa = 1$ , is weakly convergent towards  $\pi_{\gamma, \theta}^\sharp$ . Using that  $\mu \mapsto \text{KL}(\mu | \pi_\theta)$  is lower semi-continuous for any  $\theta \in \Theta$ , see [19, Lemma 1.4.3b], and [21, Corollary 18] we get that

$$\text{KL}\left(\pi_{\gamma, \theta}^\sharp | \pi_\theta\right) \leq \liminf_{n \rightarrow +\infty} \text{KL}\left(n^{-1} \sum_{k=1}^n \delta_x S_{\gamma, \theta}^k | \pi_\theta\right) \leq \gamma(Ld + M^2).$$

Using a generalized Pinsker inequality, see [22, Lemma 24], Lemma 27 and Lemma 22 if **H2** holds or Lemma 23 if **H3** holds, we get that

$$\begin{aligned} \|\pi_{\gamma, \theta}^\sharp - \pi_\theta\|_{W^{1/2}} &\leq \sqrt{2}(\pi_{\gamma, \theta}^\sharp(W) + \pi_\theta(W))^{1/2} \text{KL}\left(\pi_{\gamma, \theta}^\sharp | \pi_\theta\right)^{1/2} \\ &\leq \sqrt{2}\{b\lambda^{-\bar{\gamma}}/\log(1/\lambda) + \sup_{\theta \in \Theta} \pi_\theta(W)\}^{1/2}(Ld + M^2)^{1/2}\gamma^{1/2}, \end{aligned}$$

which concludes the proof.  $\square$

**Lemma 29.** *Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < 2/(\underline{m} + L)$  if **H2** holds and  $\bar{\gamma} < 2/L$  if **H3** holds. Then there exists  $\bar{B}_3 \geq 0$  such that for any  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $x \in \mathbb{R}^d$  and  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$  with  $i \in \{1, 2\}$  we have*

$$\max\left(\|\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} - \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\|_{W^{1/2}}, \|\delta_x \bar{S}_{1, \gamma, \theta}^{[1/\gamma]} - \delta_x \bar{S}_{2, \gamma, \theta}^{[1/\gamma]}\|_{W^{1/2}}\right) \leq \bar{B}_3 \gamma |\kappa_1 - \kappa_2| W^{1/2}(x).$$

where for any  $i \in \{1, 2\}$ ,  $\theta \in \Theta$  and  $\gamma \in (0, \bar{\gamma}]$ ,  $S_{i, \gamma, \theta}$  is given by (18) 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  $\alpha < \min(\underline{\kappa}\eta/4, \eta)$  if **H3** holds, see (19).

*Proof.* We only show that for any  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $x \in \mathbb{R}^d$  and  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$  with  $i \in \{1, 2\}$  we have  $\|\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} - \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\|_{W^{1/2}} \leq \bar{B}_3 \gamma |\kappa_1 - \kappa_2| W^{1/2}(x)$  since the proof for  $\bar{S}_{1, \gamma, \theta}$  and  $\bar{S}_{2, \gamma, \theta}$  is similar. Let  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$ ,  $x \in \mathbb{R}^d$  and  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$  with  $i \in \{1, 2\}$ . Using a generalized Pinsker inequality, see [22, Lemma 24], we have

$$\begin{aligned} \|\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} - \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\|_{W^{1/2}} &\leq \sqrt{2}(S_{1, \gamma, \theta}^{[1/\gamma]} W(x) + S_{2, \gamma, \theta}^{[1/\gamma]} W(x))^{1/2} \text{KL}\left(\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} | \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\right)^{1/2}. \end{aligned} \quad (30)$$

Using [30, Lemma 4.1] we get that  $\text{KL}\left(\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} | \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\right) \leq \text{KL}(\tilde{\mu}_1 | \tilde{\mu}_2)$  where setting  $T = \gamma \lceil 1/\gamma \rceil$ ,  $\tilde{\mu}_i$ ,  $i \in \{1, 2\}$ , is the probability measure over  $\mathcal{B}(C([0, T], \mathbb{R}^d))$  which is defined for any  $A \in \mathcal{B}(C([0, T], \mathbb{R}^d))$  by  $\tilde{\mu}_i(A) = \mathbb{P}((X_t^i)_{t \in [0, T]} \in A)$ ,  $i \in \{1, 2\}$  and for any  $t \in [0, T]$

$$dX_t^i = b_i(t, (X_s^i)_{s \in [0, T]}) dt + \sqrt{2} dB_t, \quad X_0^i = x,$$

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

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

where for any  $y \in \mathbb{R}^d$ ,  $\mathcal{T}_{\gamma, \theta}(y) = y - \gamma \nabla_x V_\theta(y)$ . Since  $(X_t^i)_{t \in [0, T]} \in C([0, T], \mathbb{R}^d)$ ,  $b_i$  and  $b$  are continuous for any  $i \in \{1, 2\}$ , [32, Theorem 7.19] applies and we obtain that  $\tilde{\mu}_1 \ll \tilde{\mu}_2$  and

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

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

$$\text{KL}(\tilde{\mu}_1 | \tilde{\mu}_2) = (1/4) \mathbb{E} \left[ \int_0^T \|b_1(t, (X_s^1)_{s \in [0, T]}) - b_2(t, (X_s^1)_{s \in [0, T]})\|^2 ds \right]. \quad (31)$$

In addition, using Lemma 11, we have for any  $(\omega_s)_{s \in [0, T]} \in C([0, T], \mathbb{R}^d)$  and  $t \in [0, T]$

$$\begin{aligned} \|b_1(t, (\omega_s)_{s \in [0, T]}) - b_2(t, (\omega_s)_{s \in [0, T]})\|^2 &= \|\mathcal{T}_{\gamma, \theta}(\text{prox}_{U_\theta}^{\gamma \kappa_1}(\omega_{\gamma \lfloor t/\gamma \rfloor})) - \mathcal{T}_{\gamma, \theta}(\text{prox}_{U_\theta}^{\gamma \kappa_2}(\omega_{\gamma \lfloor t/\gamma \rfloor}))\|^2 \\ &\leq \|\text{prox}_{U_\theta}^{\gamma \kappa_1}(\omega_{\gamma \lfloor t/\gamma \rfloor}) - \text{prox}_{U_\theta}^{\gamma \kappa_2}(\omega_{\gamma \lfloor t/\gamma \rfloor})\|^2 \leq 4\gamma^2(\kappa_1 - \kappa_2)^2 M^2. \end{aligned} \quad (32)$$

Combining this result and (31) we get that

$$KL\left(\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} | \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\right) \leq (1 + \bar{\gamma})M^2\gamma^2 |\kappa_1 - \kappa_2|^2. \quad (33)$$

Combining (33) and (30) we get that

$$\begin{aligned} \|\delta_x S_{1, \gamma, \theta}^{[1/\gamma]} - \delta_x S_{2, \gamma, \theta}^{[1/\gamma]}\|_{W^{1/2}} &\leq 2^{1/2}(1 + \bar{\gamma})^{1/2}M(S_{1, \gamma, \theta}^{[1/\gamma]} W(x) + S_{2, \gamma, \theta}^{[1/\gamma]} W(x))^{1/2}\gamma |\kappa_1 - \kappa_2|. \end{aligned}$$

We conclude the proof upon using Lemma 8, and Lemma 22 if **H2** holds, or Lemma 23 if **H3** holds.

**Proposition 30.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < 2/(\underline{m} + L)$  if **H2** holds and  $\bar{\gamma} < 2/L$  if **H3** holds. Then there exists  $B_3 \geq 0$  such that for any  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$  with  $i \in \{1, 2\}$  we have

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

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  $S_{i, \gamma, \theta}$ , respectively  $\bar{S}_{i, \gamma, \theta}$ , given by (18) 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  $\alpha < \min(\underline{\kappa}\eta/4, \eta)$  if **H3** holds, see (19).

*Proof.* We only show that for any  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$  with  $i \in \{1, 2\}$ ,  $\|\pi_{\gamma, \theta}^1 - \pi_{\gamma, \theta}^2\|_{W^{1/2}} \leq B_3\gamma |\kappa_2 - \kappa_1|$  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 25 we have

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

Let  $n = q \lceil 1/\gamma \rceil$ . Using Theorem 25 with  $a = 1/2$ , that  $W^{1/2}(x) \leq W(x)$  for any  $x \in \mathbb{R}^d$ , Lemma 29, Lemma 8 and Lemma 22 if **H2** holds or Lemma 23 if **H3** holds, we have

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

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  $x = 0$ .  $\square$

**Corollary 31.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < 2/(\underline{m} + L)$  if **H2** holds and  $\bar{\gamma} < 2/L$  if **H3** holds. Then for any  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\theta \in \Theta$  and  $\gamma \in (0, \bar{\gamma}]$ , we have

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

where for any  $\gamma \in (0, \bar{\gamma}]$ ,  $\pi_{\gamma, \theta}$  is the invariant probability measure of  $S_{\gamma, \theta}$  given by (18). In addition,  $\Psi(\gamma) = \tilde{\Psi}(\gamma) + B_3\gamma|\kappa - 1|$ , where  $\tilde{\Psi}$  is given in Theorem 28 and  $B_3$  in Proposition 30, and  $W = W_m$  with  $m \in \mathbb{N}^*$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta)$  if **H3** holds, see (19).

*Proof.* We only show that for any  $\theta \in \Theta$  and  $\gamma \in (0, \bar{\gamma}]$  we have  $\|\pi_{\gamma, \theta} - \pi_\theta\|_{W^{1/2}} \leq \Psi(\gamma)$  since the proof for  $\bar{\pi}_{\gamma, \theta}$  and  $\bar{\pi}_\theta$  are similar. Let  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$ . The proof is a direct application of Theorem 28 and Proposition 30 upon noticing that

$$\|\pi_{\gamma, \theta} - \pi_\theta\|_{W^{1/2}} \leq \|\pi_{\gamma, \theta} - \pi_{\gamma, \theta}^\sharp\|_{W^{1/2}} + \|\pi_{\gamma, \theta}^\sharp - \pi_\theta\|_{W^{1/2}},$$

where  $\pi_{\gamma, \theta}^\sharp$  is the invariant probability measure of  $S_{\gamma, \theta}$  given by (18) and associated with  $\kappa = 1$ .  $\square$

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

$$\begin{aligned} \max (\|\delta_x S_{\gamma_1, \theta_1} - \delta_x S_{\gamma_2, \theta_2}\|_{W^a}, \|\delta_x \bar{S}_{\gamma_1, \theta_1} - \delta_x \bar{S}_{\gamma_2, \theta_2}\|_{W^a}) \\ \leq (\Lambda(\gamma_1, \gamma_2) + \Lambda(\gamma_1, \gamma_2) \|\theta_1 - \theta_2\|) W^{2a}(x), \end{aligned}$$

with

$$\Lambda_1(\gamma_1, \gamma_2) = A_4(\gamma_1/\gamma_2 - 1), \quad \Lambda_2(\gamma_1, \gamma_2) = A_4 \gamma_2^{1/2},$$

and where  $W = W_m$  with  $m \in \mathbb{N}$  and  $m \geq 2$  if H2 is satisfied and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta)$  if H3 is satisfied, see (19).

*Proof.* We only show that for any  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\theta_1, \theta_2 \in \Theta$ ,  $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$  with  $\gamma_2 < \gamma_1$ ,  $a \in [1/4, 1/2]$  and  $x \in \mathbb{R}^d$  we have  $\|\delta_x S_{\gamma_1, \theta_1} - \delta_x S_{\gamma_2, \theta_2}\|_{W^a} \leq (\Lambda(\gamma_1, \gamma_2) + \Lambda(\gamma_1, \gamma_2) \|\theta_1 - \theta_2\|) W^{2a}(x)$  since the proof for  $\bar{S}_{\gamma_1, \theta_1}$  and  $\bar{S}_{\gamma_2, \theta_2}$  is similar. Let  $a \in [1/4, 1/2]$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\theta_1, \theta_2 \in \Theta$ ,  $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$  with  $\gamma_2 < \gamma_1$ . Using a generalized Pinsker inequality, see [22, Lemma 24], we have

$$\begin{aligned} \|\delta_x S_{\gamma_1, \theta_1} - \delta_x S_{\gamma_2, \theta_2}\|_{W^a} \\ \leq \sqrt{2} (\delta_x S_{\gamma_1, \theta_1} W^{2a}(x) + \delta_x S_{\gamma_2, \theta_2} W^{2a}(x))^{1/2} \text{KL}(\delta_x S_{\gamma_1, \theta_1} \| \delta_x S_{\gamma_2, \theta_2})^{1/2}. \end{aligned}$$

Combining this result, Jensen's inequality and Lemma 22 if H2 holds and Lemma 23 if H3 holds, we obtain that

$$\|S_{\gamma_1, \theta_1} - S_{\gamma_2, \theta_2}\|_{W^a} \leq 2(1 + b\bar{\gamma})^{1/2} \{\text{KL}(\delta_x S_{\gamma_1, \theta_1} \| \delta_x S_{\gamma_2, \theta_2})\}^{1/2} W^a(x).$$

Denote for  $v \in \mathbb{R}^d$  and  $\sigma > 0$ ,  $\Upsilon_{v, \sigma}$  the  $d$ -dimensional Gaussian distribution with mean  $v$  and covariance matrix  $\sigma^2 \text{Id}$ . Using Lemma 17 and the fact that  $\gamma_1 \geq \gamma_2$  we have

$$\begin{aligned} \text{KL}(\delta_x S_{\gamma_1, \theta_1} \| \delta_x S_{\gamma_2, \theta_2}) \\ \leq d(\gamma_1/\gamma_2 - 1)^2/2 + \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_2}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) \right\|^2 / (4\gamma_2), \end{aligned} \tag{34}$$

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

$$\begin{aligned} (1/4) \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_2}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\|^2 \\ \leq \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x)) - \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) \right\|^2 + \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) \right\|^2 \\ + \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\|^2 + \left\| \mathcal{T}_{\gamma_2, \theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_2}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\|^2. \end{aligned} \tag{35}$$

First using H1, [36, Theorem 2.1.5, Equation (2.1.8)] and Lemma 11 we have

$$\begin{aligned} \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x)) - \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) \right\| \\ \leq \left\| \text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x) - \text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x) \right\| \leq 2M |\gamma_1 \kappa - \gamma_2 \kappa|. \end{aligned} \tag{36}$$

Second, we have using (9), H1, [36, Theorem 2.1.5, Equation (2.1.8)] and H4

$$\begin{aligned} \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_1}(\text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x)) \right\| \\ \leq \gamma_2 \kappa \left\| \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x) - \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x) \right\| \leq \sup_{t \in [0, \bar{\gamma} \kappa]} \{f_\theta(t)\} \gamma_2 \kappa \|\theta_1 - \theta_2\| (1 + \|x\|). \end{aligned} \tag{37}$$

Third using **H1** and Lemma 9 we have that

$$\begin{aligned} \left\| \mathcal{T}_{\gamma_1, \theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\| &\leq (\gamma_1 - \gamma_2) \left\| \nabla_x V_{\theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\| \\ &\leq (\gamma_1 - \gamma_2)L \left\| \text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x) - x_{\theta_1}^* \right\| \\ &\leq (\gamma_1 - \gamma_2)L(R_{V,1} + \bar{\gamma}\kappa M + \|x\|). \end{aligned} \quad (38)$$

Finally using **H1**, **H4** and Lemma 9 we have that

$$\begin{aligned} &\left\| \mathcal{T}_{\gamma_2, \theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) - \mathcal{T}_{\gamma_2, \theta_2}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\| \\ &\leq \gamma_2 \left\| \nabla_x V_{\theta_1}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) - \nabla_x V_{\theta_2}(\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)) \right\| \\ &\leq \gamma_2 M_\Theta \|\theta_1 - \theta_2\| (1 + \|\text{prox}_{U_{\theta_2}}^{\gamma_2 \kappa}(x)\|) \leq \gamma_2 M_\Theta \|\theta_1 - \theta_2\| (1 + \bar{\gamma}\kappa M + \|x\|). \end{aligned} \quad (39)$$

Therefore, combining (36), (37), (38) and (39) in (35), there exists  $A_{4,1} \geq 0$  such that for any  $\gamma_1, \gamma_2 > 0$  with  $\gamma_2 < \gamma_1$  and  $\theta_1, \theta_2 \in \Theta$

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

Using this result in (34), there exists  $A_{4,2} \geq 0$  such that

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

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  $u, v \geq 0$ ,  $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ .  $\square$

## 5.5 Checking [17, H1, H2] for MYULA

In this section, similarly to Section 5.5 for PULA, we show that [17, H1, H2] hold for MYULA.

**Lemma 33.** Assume **H1**, **H2** or **H3**, and let  $(X_k^n, \bar{X}_k^n)_{n \in \mathbb{N}, k \in \{0, \dots, m_n\}}$  be given by (5) with  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(R_{\gamma, \theta}, \bar{R}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$  and  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  with  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Then there exists  $\bar{A}_1 \geq 1$  such that for any  $n, p \in \mathbb{N}$  and  $k \in \{0, \dots, m_n\}$

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

with  $W = W_m$  with  $m \in \mathbb{N}^*$  and  $\bar{\gamma} < 2/(\mathfrak{m} + L)$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta/8)$  and  $\bar{\gamma} < \min\{2/L, \eta/(2ML)\}$  if **H3** holds, see (19).

**Proposition 34.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/L, 2/(\mathfrak{m} + L)\}$  if **H2** holds and  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/L, \eta/(2ML)\}$  if **H3** holds. Then there exists  $\bar{B}_{3,1} \geq 0$  such that for any  $\theta \in \Theta$ ,  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma \in (0, \bar{\gamma}]$

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

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  $R_{i, \gamma, \theta}$ , respectively  $\bar{R}_{i, \gamma, \theta}$ , given by (17) 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  $\alpha < \min(\underline{\kappa}\eta/4, \eta/8)$  if **H3** holds, see (19).

*Proof.* The proof is similar to the one of Proposition 30 upon setting for any  $i \in \{1, 2\}$  and  $(\omega_s)_{s \in [0, T]} \in C([0, T], \mathbb{R}^d)$  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 (32) in Lemma 29 by

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

$\square$

**Proposition 35.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < \min\{(2 - 1/\kappa)/L, 2/(\mathfrak{m} + L), L^{-1}\}$  if **H2** holds and  $\bar{\gamma} < \min\{(2 - 1/\kappa)/L, \eta/(2ML), L^{-1}\}$  if **H3** holds. Then there exists  $\bar{B}_{3,2} \geq 0$  such that for any  $\theta \in \Theta$ ,  $\gamma \in (0, \bar{\gamma}]$  and  $\kappa_i \in [\underline{\kappa}, \bar{\kappa}]$  with  $i \in \{1, 2\}$  we have

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

where for any  $\theta \in \Theta$  and  $\gamma \in (0, \bar{\gamma}]$ ,  $\pi_{\gamma, \theta}^b$ , respectively  $\bar{\pi}_{\gamma, \theta}^b$ , is the invariant probability measure of  $R_{\gamma, \theta}$ , respectively  $\bar{R}_{\gamma, \theta}$ , given by (17) and associated with  $\kappa = 1$  and  $\pi_{\gamma, \theta}^\sharp$ , respectively  $\bar{\pi}_{\gamma, \theta}^\sharp$ , is the invariant probability measure of  $S_{\gamma, \theta}$ , respectively  $\bar{S}_{\gamma, \theta}$ , given by (18) and associated with  $\kappa = 1$ . In addition,  $W = W_m$  with  $m \in \mathbb{N}^*$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta/8)$  if **H3** holds, see (19).

*Proof.* The proof is similar to the one of Proposition 30 upon setting for any  $(\omega_s)_{s \in [0, T]} \in C([0, T], \mathbb{R}^d)$  with  $T = \gamma \lceil 1/\gamma \rceil$

$$\begin{aligned} b_1(t, (\omega_s)_{s \in [0, T]}) &= \text{prox}_{U_\theta}^\gamma(\omega_{\lfloor t/\gamma \rfloor \gamma}) - \gamma \nabla_x V_\theta(\text{prox}_{U_\theta}^\gamma(\omega_{\lfloor t/\gamma \rfloor \gamma})) , \\ b_2(t, (\omega_s)_{s \in [0, T]}) &= \omega_{\lfloor t/\gamma \rfloor \gamma} - \gamma \nabla_x V_\theta(\omega_{\lfloor t/\gamma \rfloor \gamma}) - \gamma \nabla_x U_\theta^\gamma(\omega_{\lfloor t/\gamma \rfloor \gamma}) , \end{aligned}$$

and replacing (32) in Lemma 29 and using (9) and Lemma 9 we get

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

□

**Proposition 36.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/L, 2/(\mathfrak{m} + L), L^{-1}\}$  if **H2** holds and  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/L, \eta/(2ML), L^{-1}\}$  if **H3** holds. Then for any  $\theta \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$  and  $\gamma \in (0, \bar{\gamma}]$ , we have

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

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  $R_{i, \gamma, \theta}$ , respectively  $\bar{R}_{i, \gamma, \theta}$ , given by (17) and associated with  $\kappa \leftarrow \kappa_i$ . In addition,  $\bar{\Psi}(\gamma) = \bar{\Psi}(\gamma) + \bar{B}_{3,1}\gamma + \bar{B}_{3,2}\gamma^2$ , where  $\bar{\Psi}$  is given in Theorem 28 and  $B_3$  in Proposition 30, and  $W = W_m$  with  $m \in \mathbb{N}^*$  if **H2** holds and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta/8)$  if **H3** holds, see (19).

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

$$\|\pi_{\gamma, \theta} - \pi_\theta\|_{W^{1/2}} \leq \|\pi_{\gamma, \theta} - \pi_{\gamma, \theta}^b\|_{W^{1/2}} + \|\pi_{\gamma, \theta}^b - \pi_{\gamma, \theta}^\sharp\|_{W^{1/2}} + \|\pi_{\gamma, \theta}^\sharp - \pi_\theta\|_{W^{1/2}} ,$$

where for any  $\theta \in \Theta$  and  $\gamma \in (0, \bar{\gamma}]$ ,  $\pi_{\gamma, \theta}^b$  is the invariant probability measure of  $R_{\gamma, \theta}$  given by (17) and associated with  $\kappa = 1$  and  $\pi_{\gamma, \theta}^\sharp$  is the invariant probability measure of  $S_{\gamma, \theta}$  and associated with  $\kappa = 1$ . We conclude the proof upon combining Proposition 34, Proposition 35 and Theorem 28. □

**Proposition 37.** Assume **H1** and **H2** or **H3**. Let  $\bar{\kappa} \geq 1 \geq \underline{\kappa} > 1/2$ . Let  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/L, 2/(\mathfrak{m} + L)\}$  if **H2** holds and  $\bar{\gamma} < \min\{(2 - 1/\underline{\kappa})/L, \eta/(2ML)\}$  if **H3** holds. Then there exists  $\bar{A}_4 \geq 0$  such that for any  $\theta_1, \theta_2 \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$  with  $\gamma_2 < \gamma_1$ ,  $a \in [1/4, 1/2]$  and  $x \in \mathbb{R}^d$

$$\begin{aligned} &\max (\|\delta_x R_{\gamma_1, \theta_1} - \delta_x R_{\gamma_2, \theta_2}\|_{W^a}, \|\delta_x \bar{R}_{\gamma_1, \theta_1} - \delta_x \bar{R}_{\gamma_2, \theta_2}\|_{W^a}) \\ &\leq (\bar{\Lambda}_1(\gamma_1, \gamma_2) + \bar{\Lambda}_2(\gamma_1, \gamma_2)) \|\theta_1 - \theta_2\| W^{2a}(x) , \end{aligned}$$

with

$$\bar{\Lambda}_1(\gamma_1, \gamma_2) = \bar{A}_4(\gamma_1/\gamma_2 - 1) , \quad \bar{\Lambda}_2(\gamma_1, \gamma_2) = \bar{A}_4 \gamma_2^{1/2} ,$$

and where  $W = W_m$  with  $m \in \mathbb{N}$  and  $m \geq 2$  if **H2** is satisfied and  $W = W_\alpha$  with  $\alpha < \min(\underline{\kappa}\eta/4, \eta/8)$  if **H3** is satisfied, see (19).

*Proof.* First, note that we only show that for any  $\theta_1, \theta_2 \in \Theta$ ,  $\kappa \in [\bar{\kappa}, \underline{\kappa}]$ ,  $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$  with  $\gamma_2 < \gamma_1$ ,  $a \in [1/4, 1/2]$  and  $x \in \mathbb{R}^d$ , we have  $\|\delta_x R_{\gamma_1, \theta_1} - \delta_x R_{\gamma_2, \theta_2}\|_{W^a} \leq (\bar{\Lambda}(\gamma_1, \gamma_2) + \bar{\Lambda}(\gamma_1, \gamma_2) \|\theta_1 - \theta_2\|) W^{2a}(x)$  since the proof for  $\bar{R}_{\gamma_1, \theta_1}$  and  $\bar{R}_{\gamma_2, \theta_2}$  is similar. Let  $a \in [1/4, 1/2]$ ,  $\theta_1, \theta_2 \in \Theta$ ,  $\kappa \in [\underline{\kappa}, \bar{\kappa}]$ ,  $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$  with  $\gamma_2 < \gamma_1$ . Using a generalized Pinsker inequality [22, Lemma 24] we have

$$\begin{aligned} \|\delta_x R_{\gamma_1, \theta_1} - \delta_x R_{\gamma_2, \theta_2}\|_{W^a} \\ \leq \sqrt{2}(\delta_x R_{\gamma_1, \theta_1} W^{2a}(x) + \delta_x R_{\gamma_2, \theta_2} W^{2a}(x))^{1/2} \text{KL}(\delta_x R_{\gamma_1, \theta_1} | \delta_x R_{\gamma_2, \theta_2})^{1/2}. \end{aligned}$$

Combining this result, Jensen's inequality and Lemma 22 if **H2** holds and Lemma 23 if **H3** holds, we obtain that

$$\|\delta_x R_{\gamma_1, \theta_1} - \delta_x R_{\gamma_2, \theta_2}\|_{W^a} \leq 2(1 + b\bar{\gamma})^{1/2} \text{KL}(\delta_x R_{\gamma_1, \theta_1} | \delta_x R_{\gamma_2, \theta_2})^{1/2} W^a(x).$$

Using Lemma 17 and the fact that  $\gamma_1 \geq \gamma_2$  we have

$$\begin{aligned} \text{KL}(\delta_x R_{\gamma_1, \theta_1} | \delta_x R_{\gamma_2, \theta_2}) \\ \leq d(\gamma_1/\gamma_2 - 1)^2/2 + \|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_1 \nabla_x V_{\theta_1}(x) + \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x) - \gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x)\|^2/(4\gamma_2), \end{aligned} \quad (40)$$

We have

$$\begin{aligned} \|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_1 \nabla_x V_{\theta_1}(x) + \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x) - \gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x)\|^2 \\ \leq 4 \|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_2 \nabla_x V_{\theta_1}(x)\|^2 + 4 \|\gamma_2 \nabla_x V_{\theta_1}(x) - \gamma_1 \nabla_x V_{\theta_1}(x)\|^2 \\ + 4 \|\gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x)\|^2 + 4 \|\gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x)\|^2. \end{aligned} \quad (41)$$

First using **H4** we have

$$\|\gamma_2 \nabla_x V_{\theta_2}(x) - \gamma_2 \nabla_x V_{\theta_1}(x)\| \leq \gamma_2 M_\Theta \|\theta_1 - \theta_2\| (1 + \|x\|). \quad (42)$$

Second using **H1** we have

$$\begin{aligned} \|\gamma_2 \nabla_x V_{\theta_1}(x) - \gamma_1 \nabla_x V_{\theta_1}(x)\| &\leq (\gamma_1 - \gamma_2) \|\nabla_x V_{\theta_1}(x)\| \\ &\leq (\gamma_1 - \gamma_2) L \|x - x_{\theta_1}^*\| \leq (\gamma_1 - \gamma_2) L (R_{V,1} + \|x\|). \end{aligned} \quad (43)$$

Third using **H1**, **H4**, Lemma 9 and Lemma 11 we have

$$\begin{aligned} \|\gamma_1 \nabla_x U_{\theta_1}^{\gamma_1 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x)\| &\leq \left\| (x - \text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x))/\kappa - (x - \text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x))/\kappa \right\| \\ &\leq \left\| \text{prox}_{U_{\theta_1}}^{\gamma_2 \kappa}(x) - \text{prox}_{U_{\theta_1}}^{\gamma_1 \kappa}(x) \right\| / \kappa \\ &\leq 2M(\gamma_1 - \gamma_2) \end{aligned} \quad (44)$$

Finally using **H4** we have

$$\|\gamma_2 \nabla_x U_{\theta_1}^{\gamma_2 \kappa}(x) - \gamma_2 \nabla_x U_{\theta_2}^{\gamma_2 \kappa}(x)\| \leq \gamma_2 \left\{ \sup_{[0, \bar{\gamma} \kappa]} f_\theta(t) \right\} \|\theta_1 - \theta_2\|. \quad (45)$$

Combining (42), (43), (44) and (45) in (41) we get that there exists  $\bar{A}_{4,1} \geq 0$  such that

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

Using this result in (40) we obtain that there exists  $\bar{A}_{4,2} \geq 0$  such that

$$\text{KL}(\delta_x R_{\gamma_1, \theta_1} | \delta_x R_{\gamma_2, \theta_2}) \leq \bar{A}_{4,2} [(\gamma_1/\gamma_2 - 1)^2 + \gamma_2 \|\theta_1 - \theta_2\|^2] W^{2a}(x),$$

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  $u, v \geq 0$ ,  $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ .  $\square$

## 5.6 Proof of Theorem 6

We divide the proof in two parts.

(a) First 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 (5) and we have  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(S_{\gamma, \theta}, \bar{S}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ . Then Lemma 26 implies that [17, H1a] is satisfied with  $A_1 \leftarrow A_1$ , Theorem 25 implies that [17, H1b] holds with  $A_2 \leftarrow A_2$  and  $\rho \leftarrow \rho$ . Finally, using Corollary 31 we get that [17, H1c] holds with  $\Psi \leftarrow \Psi$ . Therefore, we can apply [17, Theorem 1] and we obtain that the sequence  $(\theta_n)_{n \in \mathbb{N}}$  converges a.s. if

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

Since  $\Psi(\gamma_n) = \mathcal{O}(\gamma_n^{1/2})$  by Corollary 31, these summability conditions are satisfied under the summability assumptions of Theorem 6-(1). Proposition 32 implies that [17, H2] holds with  $\Lambda_1 \leftarrow \Lambda_1$  and  $\Lambda_2 \leftarrow \Lambda_2$ . Therefore if  $m_n = m_0$  for all  $n \in \mathbb{N}$ , we can apply [17, Theorem 3] and we obtain that the sequence  $(\theta_n)_{n \in \mathbb{N}}$  converges a.s. if

$$\begin{aligned} \sum_{n=0}^{+\infty} \delta_n &= +\infty, & \sum_{n=0}^{+\infty} \delta_{n+1} \Psi(\gamma_n) &< +\infty, & \sum_{n=0}^{+\infty} \delta_{n+1} \gamma_n^{-2} &< +\infty \\ \sum_{n=0}^{+\infty} \delta_{n+1} / \gamma_n^2 (\Lambda_1(\gamma_n, \gamma_{n+1}) + \delta_{n+1} \Lambda_2(\gamma_n, \gamma_{n+1})) &< +\infty. \end{aligned}$$

These summability conditions are satisfied under the summability assumptions of Theorem 6-(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 (5) with  $\{(K_{\gamma, \theta}, \bar{K}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{(R_{\gamma, \theta}, \bar{R}_{\gamma, \theta}) : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ . Then Lemma 33 implies that [17, H1a] is satisfied with  $A_1 \leftarrow \bar{A}_1$ , Theorem 21 implies that [17, H1b] holds with  $A_2 \leftarrow \bar{A}_2$  and  $\rho \leftarrow \bar{\rho}$ . Finally, using Proposition 36 we get that [17, H1c] holds with  $\Psi \leftarrow \bar{\Psi}$ . Therefore, we can apply [17, Theorem 1] and we obtain that the sequence  $(\theta_n)_{n \in \mathbb{N}}$  converges a.s. if

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

Since  $\bar{\Psi}(\gamma_n) = \mathcal{O}(\gamma_n^{1/2})$  by Proposition 36, these summability conditions are satisfied under the summability assumptions of Theorem 6-(1). Proposition 37 implies that [17, H2] holds with  $\Lambda_1 \leftarrow \bar{\Lambda}_1$  and  $\Lambda_2 \leftarrow \bar{\Lambda}_2$ . Therefore if  $m_n = m_0$  for all  $n \in \mathbb{N}$ , we can apply [17, Theorem 3] and we obtain that the sequence  $(\theta_n)_{n \in \mathbb{N}}$  converges a.s. if

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

These summability conditions are satisfied under the summability assumptions of Theorem 6-(2).  $\square$

## 5.7 Proof of Theorem 7

The proof is similar to the one of Theorem 6 using [16, Theorem 2, Theorem 4] instead of [16, Theorem 1, Theorem 3].

## 6 Acknowledgements

AD acknowledges financial support from Polish National Science Center grant: NCN UMO-2018/31/B/ST1/00253. MP acknowledges financial support from EPSRC under grant EP/T007346/1.

## References

- [1] Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.*, 18(1):310–342, 2017.
- [2] Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 451–459, 2011.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] M. Benaim. A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2):437–472, 1996.
- [7] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.
- [8] 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.
- [9] 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.
- [10] Emmanuel J Candès, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.
- [11] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.
- [12] 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.
- [13] Julianne Chung and Linh Nguyen. Motion estimation and correction in photoacoustic tomographic reconstruction. *SIAM Journal on Imaging Sciences*, 10(1):216–242, 2017.
- [14] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.
- [15] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.
- [16] V. De Bortoli and A. Durmus. Convergence of diffusions and their discretizations: from continuous to discrete processes and back, 2019.
- [17] 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. 2019.

- [18] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.
- [19] 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.
- [20] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *ArXiv e-prints*, May 2016.
- [21] 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.
- [22] Alain Durmus, Eric Moulines, et al. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.
- [23] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.
- [24] Bruno Galerne and Arthur Leclaire. Texture inpainting using efficient Gaussian conditional simulation. *SIAM Journal on Imaging Sciences*, 10(3):1446–1474, 2017.
- [25] 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.
- [26] Mark A Iwen, AdityaViswanathan, and YangWang. Fast phaseretrieval fromlocal correlation measurements. *SIAM Journal on Imaging Sciences*, 9(4):1655–1688, 2016.
- [27] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.
- [28] 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.
- [29] Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
- [30] Solomon Kullback. *Information theory and statistics*. John Wiley and Sons, Inc., New York, 1959.
- [31] 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.
- [32] 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.
- [33] M. Métivier and P. Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Trans. Inform. Theory*, 30(2, part 1):140–151, 1984.
- [34] M. Métivier and P. Priouret. Théorèmes de convergence presque sûre pour une classe d’algorithmes stochastiques à pas décroissant. *Probab. Theory Related Fields*, 74(3):403–428, 1987.
- [35] 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.
- [36] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [37] GeorgePólyaand 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.

- [38] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- [39] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.
- [40] 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.
- [41] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [42] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
- [43] Lorenzo Rosasco, Silvia Villa, and Bang Cong Vu. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, pages 1–27, 2019.
- [44] Carola-Bibiane Schonlieb. *Partial Differential Equation Methods for Image Inpainting*, volume 29. Cambridge University Press, 2015.
- [45] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.
- [46] Miguel Simões, Jos e 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.
- [47] Weijie Su, Stephen P. Boyd, and Emmanuel J. Cand es. A differential equation for modeling nesterov’s accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17:153:1–153:43, 2016.
- [48] V. B. Tadi  and A. Doucet. Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.*, 27(6):3255–3304, 2017.
- [49] Ana Fernandez Vidal, Valentin De Bortoli, Marcelo Pereyra, and Durmus Alain. Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach. Part I: Methodology and experiments. *SIAM Journal on Imaging Sciences*, 2020.
- [50] 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.
- [51] 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.

# Bibliography

- [1] M. V. AFONSO, J. M. BIOCAS-DIAS, AND M. A. FIGUEIREDO, *A fast algorithm for the constrained formulation of compressive image reconstruction and other linear inverse problems*, in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 4034–4037.
- [2] M. V. AFONSO, J. M. BIOCAS-DIAS, AND M. A. FIGUEIREDO, *Fast image recovery using variable splitting and constrained optimization*, IEEE Transactions on Image Processing, 19 (2010), pp. 2345–2356.
- [3] M. S. ALMEIDA AND M. A. FIGUEIREDO, *Parameter estimation for blind and non-blind deblurring using residual whiteness measures*, IEEE Transactions on Image Processing, 22 (2013), pp. 2751–2763.
- [4] S. ARRIDGE, P. MAASS, O. ÖKTEM, AND C.-B. SCHÖNLIEB, *Solving inverse problems using data-driven models*, Acta Numerica, 28 (2019), pp. 1–174.
- [5] Y. F. ATCHADE, *An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift*, Methodology and Computing in applied Probability, 8 (2006), pp. 235–254.
- [6] Y. F. ATCHADÉ, *A computational framework for empirical Bayes inference*, Statistics and Computing, 21 (2011), pp. 463–473.
- [7] Y. F. ATCHADÉ, G. FORT, AND E. MOULINES, *On perturbed proximal gradient algorithms*, J. Mach. Learn. Res, 18 (2017), pp. 310–342.
- [8] A. AZEVEDO-FILHO AND R. D. SHACHTER, *Laplace’s method approximations for probabilistic inference in belief networks with continuous variables*, in Uncertainty Proceedings 1994, Elsevier, 1994, pp. 28–36.

## BIBLIOGRAPHY

- [9] S. D. BABACAN, R. MOLINA, AND A. K. KATSAGGELOS, *Parameter estimation in TV image restoration using variational distribution approximation*, IEEE transactions on image processing, 17 (2008), pp. 326–339.
- [10] S. D. BABACAN, R. MOLINA, AND A. K. KATSAGGELOS, *Variational Bayesian super resolution*, IEEE Transactions on Image Processing, 20 (2011), pp. 984–999.
- [11] F. R. BACH AND E. MOULINES, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, in Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, 2011, pp. 451–459.
- [12] L. BALZANO, R. NOWAK, AND J. ELLENBERG, *Compressed sensing audio demonstration*, website <http://web.eecs.umich.edu/~girasole/csaudio>, (2010).
- [13] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, Cham, second ed., 2017, <https://doi.org/10.1007/978-3-319-48311-5>, <https://doi.org/10.1007/978-3-319-48311-5>. With a foreword by Hédy Attouch.
- [14] F. BENVENUTO AND C. CAMPI, *A discrepancy principle for the landweber iteration based on risk minimization*, Applied Mathematics Letters, 96 (2019), pp. 1 – 6, <https://doi.org/https://doi.org/10.1016/j.aml.2019.04.005>, <http://www.sciencedirect.com/science/article/pii/S089396591930151X>.
- [15] J. M. BIOUCAS-DIAS, A. PLAZA, N. DOBIGEON, M. PARENTE, Q. DU, P. GADER, AND J. CHANUSSOT, *Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches*, IEEE journal of selected topics in applied earth observations and remote sensing, 5 (2012), pp. 354–379.
- [16] S. BOBKOV AND M. MADIMAN, *Concentration of the information in data with log-concave distributions*, Annals of Probability, 39 (2011),

- pp. 1528–1543, <https://doi.org/10.1214/10-AOP592>, <https://arxiv.org/abs/1012.5457>.
- [17] L. BOTTOU, *Stochastic gradient descent tricks*, in Neural networks: Tricks of the trade, Springer, 2012, pp. 421–436.
  - [18] K. BREDIES, K. KUNISCH, AND T. POCK, *Total generalized variation*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 492–526.
  - [19] N. BROSSE, A. DURMUS, É. MOULINES, AND S. SABANIS, *The tamed unadjusted Langevin algorithm*, Stochastic Processes and their Applications, (2018).
  - [20] X. CAI, M. PEREYRA, AND J. D. MC EWEN, *Uncertainty quantification for radio interferometric imaging—i. proximal mcmc methods*, Monthly Notices of the Royal Astronomical Society, 480 (2018), pp. 4154–4169.
  - [21] X. CAI, M. PEREYRA, AND J. D. MC EWEN, *Quantifying uncertainty in high dimensional inverse problems by convex optimisation*, in 2019 27th European Signal Processing Conference (EUSIPCO), IEEE, 2019, pp. 1–5.
  - [22] L. CALATRONI, C. CAO, J. C. DE LOS REYES, C.-B. SCHÖNLIEB, AND T. VALKONEN, *Bilevel approaches for learning of variational imaging models*, Variational Methods: In Imaging and Geometric Control, 18 (2017), p. 2.
  - [23] D. CALVETTI, S. MORIGI, L. REICHEL, AND F. SGALLARI, *Tikhonov regularization and the l-curve for large discrete ill-posed problems*, Journal of computational and applied mathematics, 123 (2000), pp. 423–446.
  - [24] D. CALVETTI, M. PRAGLIOLA, E. SOMERSALO, AND A. STRANG, *Sparse reconstructions from few noisy data: analysis of hierarchical bayesian models with generalized gamma hyperpriors*, Inverse Problems, 36 (2020), p. 025010.
  - [25] D. CALVETTI, E. SOMERSALO, AND A. STRANG, *Hierachical bayesian models and sparsity: ℓ 2-magic*, Inverse Problems, 35 (2019), p. 035003.
  - [26] E. J. CANDES, Y. C. ELDAR, T. STROHMER, AND V. VORONINSKI, *Phase retrieval via matrix completion*, SIAM review, 57 (2015), pp. 225–251.

## BIBLIOGRAPHY

- [27] E. J. CANDÈS ET AL., *Compressive sampling*, in Proceedings of the international congress of mathematicians, vol. 3, Madrid, Spain, 2006, pp. 1433–1452.
- [28] E. J. CANDES, C. A. SING-LONG, AND J. D. TRZASKO, *Unbiased risk estimates for singular value thresholding and spectral estimators*, IEEE transactions on signal processing, 61 (2013), pp. 4643–4657.
- [29] E. J. CANDÈS AND M. B. WAKIN, *An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]*, IEEE signal processing magazine, 25 (2008), pp. 21–30.
- [30] B. P. CARLIN AND T. A. LOUIS, *Empirical Bayes: past, present and future*, J. Amer. Statist. Assoc., 95 (2000), pp. 1286–1289, <https://doi.org/10.2307/2669771>, <https://doi.org/10.2307/2669771>.
- [31] R. E. CARRILLO, J. MC EWEN, AND Y. WIAUX, *Sparsity Averaging Reweighted Analysis (SARA): a novel algorithm for radio-interferometric imaging*, Monthly Notices of the Royal Astronomical Society, 426 (2012), pp. 1223–1234.
- [32] R. E. CARRILLO, J. D. MC EWEN, AND Y. WIAUX, *Sparsity averaging reweighted analysis (sara): a novel algorithm for radio-interferometric imaging*, Monthly Notices of the Royal Astronomical Society, 426 (2012), pp. 1223–1234.
- [33] G. CASELLA, *Empirical bayes gibbs sampling*, Biostatistics, 2 (2001), pp. 485–500.
- [34] A. CHAMBOLLE AND P.-L. LIONS, *Image recovery via total variation minimization and related problems*, Numerische Mathematik, 76 (1997), pp. 167–188.
- [35] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of mathematical imaging and vision, 40 (2011), pp. 120–145.
- [36] A. CHAMBOLLE AND T. POCK, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.

- [37] G. CHANTAS, N. GALATSANOS, A. LIKAS, AND M. SAUNDERS, *Variational bayesian image restoration based on a product of t-distributions image prior*, IEEE transactions on image processing, 17 (2008), pp. 1795–1805.
- [38] G. CHANTAS, N. P. GALATSANOS, R. MOLINA, AND A. K. KATSAGGELOS, *Variational bayesian image restoration with a product of spatially weighted total variation image priors*, IEEE transactions on image processing, 19 (2009), pp. 351–362.
- [39] E. CHOUZENOUX, A. JEZIERSKA, J.-C. PESQUET, AND H. TALBOT, *A Convex Approach for Image Restoration with Exact Poisson–Gaussian Likelihood*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 2662–2682.
- [40] J. CHUNG AND L. NGUYEN, *Motion estimation and correction in photoacoustic tomographic reconstruction*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 216–242.
- [41] M. A. CLYDE, J. O. BERGER, F. BULLARD, E. B. FORD, W. H. JEFFERYS, AND R. LUO, *Current challenges in Bayesian model choice*, Statistical Challenges in Modern Astronomy IV ASP Conference Series, 371 (2007), pp. 224–240.
- [42] P. L. COMBETTES AND J.-C. PESQUET, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, Springer, 2011, pp. 185–212.
- [43] L. CONDAT, *Matlab code for total generalized variation denoising*, 2016, <https://www.gipsa-lab.grenoble-inp.fr/~laurent.condat/download/TGVdenoise.m> (accessed 2018-06-24).
- [44] M. CONGEDO, C. GOUY-PAILLER, AND C. JUTTEN, *On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics*, Clinical Neurophysiology, 119 (2008), pp. 2677–2686.
- [45] V. DE BORTOLI, D. ALAIN, M. PEREYRA, AND A. F. VIDAL, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse*

- problems: an empirical bayesian approach. Part II: Theoretical analysis*, SIAM Journal on Imaging Sciences, (2020), p. to appear.
- [46] 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*, (2019), <https://arxiv.org/abs/1906.12281>.
- [47] V. DE BORTOLI, A. DURMUS, A. F. VIDAL, AND M. PEREYRA, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach. Part II: Theoretical analysis*, (2020), <https://arxiv.org/abs/2008.05793>.
- [48] J. C. DE LOS REYES, C.-B. SCHÖNLIEB, AND T. VALKONEN, *Bilevel parameter learning for higher-order total variation regularisation models*, Journal of Mathematical Imaging and Vision, 57 (2017), pp. 1–25.
- [49] C.-A. DELEDALLE, S. VAITER, J. FADILI, AND G. PEYRÉ, *Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 2448–2487.
- [50] B. DELYON, M. LAVIELLE, E. MOULINES, ET AL., *Convergence of a stochastic approximation version of the EM algorithm*, The Annals of Statistics, 27 (1999), pp. 94–128.
- [51] A. P. DEMPSTER, N. M. LAIRD, AND D. B. RUBIN, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), 39 (1977), pp. 1–22.
- [52] D. L. DONOHO, *De-noising by soft-thresholding*, IEEE transactions on information theory, 41 (1995), pp. 613–627.
- [53] D. L. DONOHO, *Compressed sensing*, IEEE Transactions on information theory, 52 (2006), pp. 1289–1306.
- [54] R. DOUC, E. MOULINES, AND D. STOFFER, *Nonlinear time series: Theory, methods and applications with R examples*, Chapman and Hall/CRC, 2014.

## BIBLIOGRAPHY

- [55] J. DUAN, *An introduction to stochastic dynamics*, vol. 51, Cambridge University Press, 2015.
- [56] A. DURMUS, E. MOULINES, AND M. PEREYRA, *Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 473–506.
- [57] Y. C. ELDAR, *Generalized SURE for exponential families: Applications to regularization*, IEEE Transactions on Signal Processing, 57 (2009), pp. 471–481.
- [58] H. W. ENGL, M. HANKE, AND A. NEUBAUER, *Regularization of inverse problems*, vol. 375, Springer Science & Business Media, 1996.
- [59] M. A. FIGUEIREDO, R. D. NOWAK, AND S. J. WRIGHT, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE Journal of selected topics in signal processing, 1 (2007), pp. 586–597.
- [60] G. FORT, E. OLLIER, AND A. SAMSON, *Stochastic proximal-gradient algorithms for penalized mixed models*, Statistics and Computing, 29 (2019), pp. 231–253.
- [61] N. FRIEL AND J. WYSE, *Estimating the evidence – a review*, Statistica Neerlandica, (2012), <https://doi.org/DOI:10.1111/j.1467-9574.2011.00515.x>.
- [62] B. GALERNE AND A. LECLAIRE, *Texture inpainting using efficient Gaussian conditional simulation*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 1446–1474.
- [63] J. E. GENTLE, W. K. HÄRDLE, AND Y. MORI, *Handbook of computational statistics: concepts and methods*, Springer Science & Business Media, 2012.
- [64] R. Giry, M. Elad, AND Y. C. Eldar, *The projected GSURE for automatic parameter tuning in iterative shrinkage methods*, Applied and Computational Harmonic Analysis, 30 (2011), pp. 407–422.

## BIBLIOGRAPHY

- [65] G. H. GOLUB, M. HEATH, AND G. WAHBA, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.
- [66] P. J. GREEN, K. LATUŚZYSKI, M. PEREYRA, AND C. P. ROBERT, *Bayesian computation: a summary of the current state, and samples backwards and forwards*, Statistics and Computing, 25 (2015), pp. 835–862.
- [67] J. HADAMARD, *Sur les problèmes aux dérivées partielles et leur signification physique*, Princeton university bulletin, (1902), pp. 49–52.
- [68] P. C. HANSEN, *The l-curve and its use in the numerical treatment of inverse problems*, (1999).
- [69] P. C. HANSEN AND D. P. O’LEARY, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM Journal on Scientific Computing, 14 (1993), pp. 1487–1503.
- [70] P. F. HARRISON, *Image Texture Tools: Texture Synthesis, Texture Transfer, and Plausible Restoration*, PhD thesis, 2005.
- [71] B. HARROUÉ, J.-F. GIOVANNELLI, AND M. PEREYRA, *Sélection de modèles en restauration d’image: approche bayésienne dans le cas gaussien*, in GRESTI, 2019. hal-02493284.
- [72] F. HAWARY, G. BOISSONI, C. GUILLEMOT, AND P. GUILLOTEL, *Compressive 4d light field reconstruction using orthogonal frequency selection*, in 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3863–3867.
- [73] T. Q. HUY, H. H. TUE, T. T. LONG, AND T. DUC-TAN, *Deterministic compressive sampling for high-quality image reconstruction of ultrasound tomography*, BMC medical imaging, 17 (2017), pp. 1–16.
- [74] N. IKEDA AND S. WATANABE, *Stochastic differential equations and diffusion processes*, vol. 24 of North-Holland Mathematical Library, North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second ed., 1989.

## BIBLIOGRAPHY

- [75] M.-D. IORDACHE, J. M. BIOUCAS-DIAS, AND A. PLAZA, *Total variation spatial regularization for sparse hyperspectral unmixing*, IEEE Transactions on Geoscience and Remote Sensing, 50 (2012), pp. 4484–4502.
- [76] M. A. IWEN, A. VISWANATHAN, AND Y. WANG, *Fast phase retrieval from local correlation measurements*, SIAM Journal on Imaging Sciences, 9 (2016), pp. 1655–1688.
- [77] M. I. JORDAN, Z. GHAHRAMANI, T. S. JAAKKOLA, AND L. K. SAUL, *An introduction to variational methods for graphical models*, Machine learning, 37 (1999), pp. 183–233.
- [78] J. KAIPIO AND E. SOMERSALO, *Statistical and computational inverse problems*, vol. 160, Springer Science & Business Media, 2006.
- [79] K. KAMARY, K. MENGERSEN, C. P. ROBERT, AND J. ROUSSEAU, *Testing hypotheses via a mixture estimation model*, arXiv preprint arXiv:1412.2044, (2014).
- [80] M. KECH AND F. KRAHMER, *Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems*, SIAM Journal on Applied Algebra and Geometry, 1 (2017), pp. 20–37.
- [81] S.-J. KIM, K. KOH, M. LUSTIG, S. BOYD, AND D. GORINEVSKY, *A method for large-scale  $l_1$ -regularized least squares*, IEEE Journal on Selected Topics in Signal Processing, 1 (2007), pp. 606–617.
- [82] B. KNAPIK, B. SZABÓ, A. VAN DER VAART, AND J. VAN ZANTEN, *Bayes procedures for adaptive inference in inverse problems for the white noise model*, Probability Theory and Related Fields, 164 (2016), pp. 771–813.
- [83] K. KUNISCH AND T. POCK, *A bilevel optimization approach for parameter learning in variational models*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 938–983.
- [84] H. KUSHNER AND G. G. YIN, *Stochastic approximation and recursive algorithms and applications*, vol. 35, Springer Science & Business Media, 2003.

## BIBLIOGRAPHY

- [85] K. LANGE, *A gradient algorithm locally equivalent to the em algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), 57 (1995), pp. 425–437.
- [86] A. LANZA, S. MORIGI, M. PRAGLIOLA, AND F. SGALLARI, *Space-variant tv regularization for image restoration*, in European Congress on Computational Methods in Applied Sciences and Engineering, Springer, 2017, pp. 160–169.
- [87] A. LANZA, S. MORIGI, F. SGALLARI, AND A. J. YEZZI, *Variational image denoising based on autocorrelation whiteness*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1931–1955.
- [88] C. L. LAWSON AND R. J. HANSON, *Solving least squares problems*, vol. 15, Siam, 1995.
- [89] M. LEBRUN, A. BUADES, AND J. MOREL, *A nonlocal bayesian image denoising algorithm*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1665–1688, <https://doi.org/10.1137/120874989>, <https://doi.org/10.1137/120874989>, <https://arxiv.org/abs/https://doi.org/10.1137/120874989>.
- [90] S. LI, X. KANG, L. FANG, J. HU, AND H. YIN, *Pixel-level image fusion: A survey of the state of the art*, Information Fusion, 33 (2017), pp. 100–112.
- [91] S. G. LINGALA AND M. JACOB, *A blind compressive sensing frame work for accelerated dynamic mri*, in 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 2012, pp. 1060–1063.
- [92] A. LUCAS, M. ILIADIS, R. MOLINA, AND A. K. KATSAGGELOS, *Using deep neural networks for inverse problems in imaging: beyond analytical methods*, IEEE Signal Processing Magazine, 35 (2018), pp. 20–36.
- [93] F. LUCKA, K. PROKSCH, C. BRUNE, N. BISSANTZ, M. BURGER, H. DETTE, AND F. WÜBBELING, *Risk estimators for choosing regularization parameters in ill-posed problems-properties and limitations*, Inverse Problems & Imaging, 12 (2018), pp. 1121–1155.

## BIBLIOGRAPHY

- [94] Y. MARNISSI, Y. ZHENG, E. CHOUZENOUX, AND J.-C. PESQUET, *A Variational Bayesian Approach for Image Restoration?Application to Image Deblurring With Poisson–Gaussian Noise*, IEEE Transactions on Computational Imaging, 3 (2017), pp. 722–737.
- [95] J. MARROQUIN, S. MITTER, AND T. POGGIO, *Probabilistic solution of ill-posed problems in computational vision*, Journal of the american statistical association, 82 (1987), pp. 76–89.
- [96] A. MOHAMMAD-DJAFARI, *A full Bayesian approach for inverse problems*, in Maximum entropy and Bayesian methods, Springer, 1996, pp. 135–144.
- [97] A. MOHAMMAD-DJAFARI, *Joint estimation of parameters and hyperparameters in a Bayesian approach of solving inverse problems*, in Proceedings of 3rd IEEE International Conference on Image Processing, vol. 2, IEEE, 1996, pp. 473–476.
- [98] R. MOLINA, A. K. KATSAGGELOS, AND J. MATEOS, *Bayesian and regularization methods for hyperparameter estimation in image restoration*, IEEE Transactions on Image Processing, 8 (1999), pp. 231–246.
- [99] V. MONGA, *Handbook of Convex Optimization Methods in Imaging Science*, Springer, 2017.
- [100] V. I. MORGENSHTERN AND E. J. CANDES, *Super-resolution of positive sources: The discrete setup*, SIAM Journal on Imaging Sciences, 9 (2016), pp. 412–444.
- [101] V. A. MOROZOV, *Methods for solving incorrectly posed problems*, Springer Science & Business Media, 2012.
- [102] R. M. NEAL, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media, 2012.
- [103] M. PEREYRA, *Proximal markov chain monte carlo algorithms*, Statistics and Computing, 26 (2016), pp. 745–760.

- [104] M. PEREYRA, *Revisiting Maximum-A-Posteriori Estimation in Log-Concave Models*, SIAM Journal on Imaging Sciences, 12 (2019), pp. 650–670.
- [105] M. PEREYRA, J. M. BIOCAS-DIAS, AND M. A. FIGUEIREDO, *Maximum-a-posteriori estimation with unknown regularisation parameters*, in Signal Processing Conference (EUSIPCO), 2015 23rd European, IEEE, 2015, pp. 230–234.
- [106] M. PEREYRA, N. DOBIGEON, H. BATATIA, AND J.-Y. TOURNERET, *Estimating the granularity coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm*, IEEE Transactions on Image Processing, 22 (2013), pp. 2385–2397.
- [107] M. PEREYRA, N. DOBIGEON, H. BATATIA, AND J.-Y. TOURNERET, *Computing the Cramer–Rao bound of Markov random field parameters: application to the Ising and the Potts models*, IEEE Signal Processing Letters, 21 (2014), pp. 47–50.
- [108] M. PEREYRA, P. SCHNITER, E. CHOUZENOUX, J.-C. PESQUET, J.-Y. TOURNERET, A. O. HERO, AND S. McLAUGHLIN, *A survey of stochastic simulation and optimization methods in signal processing*, IEEE Journal of Selected Topics in Signal Processing, 10 (2016), pp. 224–241.
- [109] M. PEREYRA, N. WHITELEY, C. ANDRIEU, AND J.-Y. TOURNERET, *Maximum marginal likelihood estimation of the granularity coefficient of a Potts-Markov random field within an mcmc algorithm*, in Statistical Signal Processing (SSP), 2014 IEEE Workshop on, IEEE, 2014, pp. 121–124.
- [110] J.-C. PESQUET, A. BENAZZA-BENYAHIA, AND C. CHAUX, *A SURE approach for digital signal/image deconvolution problems*, IEEE Transactions on Signal Processing, 57 (2009), pp. 4616–4632.
- [111] S. PETRONE, J. ROUSSEAU, AND C. SCRICCIOLI, *Bayes and empirical Bayes: do they merge?*, Biometrika, 101 (2014), pp. 285–302, <https://doi.org/10.1093/biomet/ast067>, <https://doi.org/10.1093/biomet/ast067>.

## BIBLIOGRAPHY

- [112] N. G. POLSON, J. G. SCOTT, AND J. WINDLE, *Bayesian inference for logistic models using pólya–gamma latent variables*, Journal of the American statistical Association, 108 (2013), pp. 1339–1349.
- [113] A. RAKHLIN, O. SHAMIR, AND K. SRIDHARAN, *Making gradient descent optimal for strongly convex stochastic optimization*, arXiv preprint arXiv:1109.5647, (2011).
- [114] S. RAMANI, T. BLU, AND M. UNSER, *Monte-Carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms*, IEEE Transactions on Image Processing, 17 (2008), pp. 1540–1554.
- [115] S. RAVISHANKAR AND Y. BRESLER, *Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 2519–2557.
- [116] A. REPETTI, M. PEREYRA, AND Y. WIAUX, *Scalable bayesian uncertainty quantification in imaging inverse problems via convex optimization*, SIAM Journal on Imaging Sciences, 12 (2019), pp. 87–118.
- [117] H. ROBBINS, *An empirical Bayes approach to statistics*, in Herbert Robbins Selected Papers, Springer, 1985, pp. 41–47.
- [118] C. ROBERT, *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media, 2007.
- [119] C. P. ROBERT AND G. CASELLA, *Monte Carlo Statistical Methods (2nd ed.)*, Springer-Verlag, New York, 2004.
- [120] C. P. ROBERT AND D. WRAITH, *Computational methods for bayesian model choice*, in Aip conference proceedings, vol. 1193, AIP, 2009, pp. 251–262.
- [121] G. O. ROBERTS AND R. L. TWEEDIE, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363, <https://doi.org/10.2307/3318418>.

## BIBLIOGRAPHY

- [122] Y. ROMANO, M. ELAD, AND P. MILANFAR, *The little engine that could: Regularization by denoising (red)*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 1804–1844.
- [123] J. ROUSSEAU AND B. SZABO, *Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator*, Ann. Statist., 45 (2017), pp. 833–865, <https://doi.org/10.1214/16-AOS1469>, <https://doi.org/10.1214/16-AOS1469>.
- [124] T. SANDERS, R. B. PLATTE, AND R. D. SKEEL, *Effective new methods for automated parameter selection in regularized inverse problems*, Applied Numerical Mathematics, (2020).
- [125] C.-B. SCHÖNLIEB, *Partial Differential Equation Methods for Image Inpainting*, vol. 29, Cambridge University Press, 2015.
- [126] M. SIMÕES, J. BIOUCAS-DIAS, L. B. ALMEIDA, AND J. CHANUSSOT, *A convex formulation for hyperspectral image superresolution via subspace-based regularization*, IEEE Transactions on Geoscience and Remote Sensing, 53 (2015), pp. 3373–3388.
- [127] C. M. STEIN, *Estimation of the mean of a multivariate normal distribution*, The annals of Statistics, (1981), pp. 1135–1151.
- [128] A. M. THOMPSON, J. C. BROWN, J. W. KAY, AND D. M. TITTERINGTON, *A study of methods of choosing the smoothing parameter in image restoration by regularization*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13 (1991), pp. 326–339.
- [129] M. E. TIPPING, *Sparse bayesian learning and the relevance vector machine*, Journal of machine learning research, 1 (2001), pp. 211–244.
- [130] M. E. TIPPING AND A. C. FAUL, *Fast marginal likelihood maximisation for sparse Bayesian models*, in Proceedings of the ninth international workshop on artificial intelligence and statistics, , Jan 2003.

## BIBLIOGRAPHY

- [131] C. VAN CHUNG, J. DE LOS REYES, AND C. SCHÖNLIEB, *Learning optimal spatially-dependent regularization parameters in total variation image denoising*, Inverse Problems, 33 (2017), p. 074005.
- [132] A. W. VAN DER VAART, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.
- [133] J. M. VARAH, *Pitfalls in the numerical solution of linear ill-posed problems*, SIAM Journal on Scientific and Statistical Computing, 4 (1983), pp. 164–176.
- [134] L. VARGAS, M. PEREYRA, AND K. C. ZYGALAKIS, *Accelerating proximal Markov chain Monte Carlo by using explicit stabilised methods*, arXiv e-prints, (2019), arXiv:1908.08845, p. arXiv:1908.08845, <https://arxiv.org/abs/1908.08845>.
- [135] S. V. VENKATAKRISHNAN, C. A. BOUMAN, AND B. WOHLBERG, *Plug-and-play priors for model based reconstruction*, in 2013 IEEE Global Conference on Signal and Information Processing, IEEE, 2013, pp. 945–948.
- [136] A. F. VIDAL, V. DE BORTOLI, M. PEREYRA, AND D. ALAIN, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach. Part I: Setting and experiments.*, SIAM Journal on Imaging Sciences, (2020), p. to appear.
- [137] A. F. VIDAL, V. DE BORTOLI, M. PEREYRA, AND A. DURMUS, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach. Part I: Methodology and experiments*, (2019), <https://arxiv.org/abs/1911.11709>.
- [138] A. F. VIDAL AND M. PEREYRA, *Maximum likelihood estimation of regularisation parameters*, in 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 1742–1746.
- [139] J. WAKEFIELD, *Bayesian and frequentist regression methods*, Springer Science & Business Media, 2013.
- [140] S. WATANABE, *A widely applicable bayesian information criterion*, Journal of Machine Learning Research, 14 (2013), pp. 867–897.

## BIBLIOGRAPHY

- [141] G. C. WEI AND M. A. TANNER, *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms*, Journal of the American statistical Association, 85 (1990), pp. 699–704.
- [142] D. S. WELLER, S. RAMANI, J.-F. NIELSEN, AND J. A. FESSLER, *Monte Carlo SURE-based parameter selection for parallel magnetic resonance imaging reconstruction*, Magnetic resonance in medicine, 71 (2014), pp. 1760–1770.
- [143] H. WHITE, *Maximum Likelihood Estimation of Misspecified Models*, Econometrica, 50 (1982), pp. 1–25.
- [144] X. YUAN AND R. HAIMI-COHEN, *Image compression based on compressive sensing: End-to-end comparison with jpeg*, IEEE Transactions on Multimedia, (2020).
- [145] Y. ZHANG, Y. TIAN, Y. KONG, B. ZHONG, AND Y. FU, *Residual dense network for image super-resolution*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.
- [146] M. V. ZIBETTI, F. S. BAZÁN, AND J. MAYER, *Determining the regularization parameters for super-resolution problems*, Signal Processing, 88 (2008), pp. 2890–2901.