



**HAL**  
open science

# Asymptotically Exact Data Augmentation: Models, Properties, and Algorithms

Maxime Vono, Nicolas Dobigeon, Pierre Chainais

► **To cite this version:**

Maxime Vono, Nicolas Dobigeon, Pierre Chainais. Asymptotically Exact Data Augmentation: Models, Properties, and Algorithms. *Journal of Computational and Graphical Statistics*, 2021, 30 (2), pp.335-348. 10.1080/10618600.2020.1826954 . hal-03064884

**HAL Id: hal-03064884**

**<https://hal.science/hal-03064884v1>**

Submitted on 26 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Asymptotically exact data augmentation: models, properties and algorithms

Maxime Vono and Nicolas Dobigeon

Univ. of Toulouse, IRIT/INP-ENSEEIH, Toulouse, France

and

Pierre Chainais

Univ. of Lille, Centrale Lille, UMR CNRS 9189 - CRISAL, Lille, France

September 14, 2020

## Abstract

Data augmentation, by the introduction of auxiliary variables, has become an ubiquitous technique to improve convergence properties, simplify the implementation or reduce the computational time of inference methods such as Markov chain Monte Carlo ones. Nonetheless, introducing appropriate auxiliary variables while preserving the initial target probability distribution and offering a computationally efficient inference cannot be conducted in a systematic way. To deal with such issues, this paper studies a unified framework, coined asymptotically exact data augmentation (AXDA), which encompasses both well-established and more recent approximate augmented models. In a broader perspective, this paper shows that AXDA models can benefit from interesting statistical properties and yield efficient inference algorithms. In non-asymptotic settings, the quality of the proposed approximation is assessed with several theoretical results. The latter are illustrated on standard statistical problems. Supplementary materials including computer code for this paper are available online.

*Keywords:* Approximation, auxiliary variables, divide-and-conquer, Bayesian inference, robustness.

# 1 Introduction

Starting at least from the 1960s with the seminal paper of [Hartley \(1958\)](#) on the expectation-maximization (EM) algorithm, introducing auxiliary variables has been a widely adopted strategy to derive iterative algorithms able to deal with possibly complicated inference problems. Indeed, either by coming from statistical physics ([Swendsen and Wang 1987](#)) or by the broad statistical community ([Dempster et al. 1977](#)), auxiliary (also called latent) variables have been used to improve ([Duane et al. 1987](#); [Edwards and Sokal 1988](#); [Marnissi et al. 2018](#)) and/or simplify ([Tanner and Wong 1987](#); [Doucet et al. 2002](#)) inference methods, such as maximum likelihood (ML) estimation or simulation-based ones. Insightful reviews of these methods were conducted by [Besag and Green \(1993\)](#); [van Dyk and Meng \(2001\)](#); [Tanner and Wong \(2010\)](#). Among many others, slice sampling and half-quadratic (HQ) methods are archetypal instances of such auxiliary variable-based methods. These methods, by introducing auxiliary variables, appear to be an interesting alternative when sampling cannot be performed directly from a target distribution  $\pi$ . Nonetheless, the superiority of simulation-based algorithms based on data augmentation (DA) over classical Markov chain Monte Carlo (MCMC) methods without DA is not obvious as pointed out by [Polson \(1996\)](#); [Damien et al. \(1999\)](#). DA methods have been found to be slower than single-site update approaches in some cases ([Hurn 1997](#)) and some improvements have been derived to cope with these problems such as partial decoupling ([Higdon 1998](#)) or the introduction of a working parameter ([Meng and van Dyk 1997](#)). Moreover, DA techniques are often used on a case-by-case basis ([Geman and Reynolds 1992](#); [Albert and Chib 1993](#); [Geman and Yang 1995](#); [Polson et al. 2013](#)) and could not be applied in general scenarios due to the absence of exact DA schemes yielding an efficient inference and low computation costs.

Similarly to approximate Bayesian computation (ABC) methods to circumvent intractable likelihoods ([Beaumont et al. 2002](#); [Sisson et al. 2018b](#)), these limitations can be tackled by considering approximate DA schemes that become exact asymptotically. For instance, inspired from the variable splitting technique used in the alternating direction method of multipliers (ADMM) ([Boyd et al. 2011](#)), [Vono et al. \(2019\)](#) and [Rendell et al. \(2020\)](#) recently and independently proposed a novel and broad Bayesian inference framework that can circumvent limitations of exact DA approaches. By introducing a collection

of instrumental (also called *splitting*) variables, the aforementioned authors considered the inference from an approximate probability distribution which can be simpler, more efficient and distributed over multiple computational workers (e.g., machines or kernels).

This paper aims at deeply investigating a broad framework coined asymptotically exact data augmentation (AXDA) which encompasses previously proposed special instances such as approximate models used in [Vono et al. \(2019\)](#); [Rendell et al. \(2020\)](#), among others. More precisely, Section 2 details how such models can be built in a quasi-systematic and simple way which is highly appreciable compared to the case-by-case search of computationally efficient DA schemes. In Section 3, we revisit some already-proposed special instances of AXDA models in order to show the potential benefits of AXDA on specific examples and to exhibit interesting properties which can be generally inherited by AXDA approaches. In Section 4, we assess quantitatively the bias of AXDA models with non-asymptotic theoretical results by considering Wasserstein and total variation distances. Then, Section 5 illustrates the previous theoretical results and the benefits of the proposed methodology on several statistical problems. In order to facilitate the use of AXDA, we eventually point out that the supplementary material involves a dedicated section (Section 5) presenting how such models can be instantiated to perform efficient inference through classical simulation-based, variational Bayes (VB), optimization or expectation-maximization (EM) methods. The proofs are also given in the supplementary material, see Section 1.

## 2 Asymptotically exact data augmentation

This section introduces AXDA schemes that aim to circumvent exact DA main issue: *the art* ([van Dyk and Meng 2001](#)) of finding the exact DA associated to a statistical model and its inference limitations. For sake of simplicity, with little abuse, we shall use the same notations for a probability distribution and its associated probability density function (pdf).

### 2.1 Motivations

In this paper, we are interested in performing the inference of a variable of interest  $\theta \in \Theta \subseteq \mathbb{R}^d$ , where  $\Theta$  is a closed convex set and  $\dim(\Theta) = d$ , by relying on a probability distribution

with density  $\pi$  writing

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta})), \quad \text{or} \quad \pi(\mathbf{y}|\boldsymbol{\theta}) \propto \exp(-f(y; \boldsymbol{\theta})), \quad (1)$$

where the potential  $f$  taking values in  $\mathbb{R}$  is such that  $\pi$  defines a proper, bounded and continuous probability distribution. For sake of generality, note that  $\pi$  in (1) shall describe various quantities. First, with a little abuse of notations,  $\pi(\boldsymbol{\theta})$  may simply refer to a pdf associated to the random variable  $\boldsymbol{\theta}$ , e.g., its prior distribution  $\pi(\boldsymbol{\theta})$  or its posterior distribution  $\pi(\boldsymbol{\theta}) \triangleq \pi(\boldsymbol{\theta}|\mathbf{y})$  when referring to a set of observations denoted by  $\mathbf{y}$ . Depending on the problem, we also allow  $\pi$  to stand for a likelihood function  $\pi(\mathbf{y}|\boldsymbol{\theta})$ . We will work under this convention and write explicitly the form of  $\pi$  when required. For sake of simplicity and clarity, only the case corresponding to  $\pi(\boldsymbol{\theta})$  will be detailed in this section. The application of the proposed methodology to  $\pi(\mathbf{y}|\boldsymbol{\theta})$  is very similar and can be retrieved by a straightforward derivation.

We consider situations where direct inference from (1) is difficult because intractable or computationally prohibitive. To overcome these issues, an option is to rely on exact DA which introduces some auxiliary variables stacked into a vector  $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^k$  and defines a new density, simpler to handle, such that

$$\int_{\mathcal{Z}} \pi(\boldsymbol{\theta}, \mathbf{z}) d\mathbf{z} = \pi(\boldsymbol{\theta}). \quad (2)$$

Much research has been devoted to these models in order to simplify an inference task or to improve the convergence properties of direct inference approaches (e.g., slice sampling and HQ methods introduced in Section 1). Nonetheless, these approaches have several limitations. Indeed, finding a convenient form for the augmented density in order to satisfy (2) while leading to efficient algorithms generally requires some knowledge and can even be impossible in some cases (Geman and Yang 1995). For instance, the mixture representation of a binomial likelihood function based on the Poly-Gamma distribution has been used to derive a promising Gibbs sampler for logistic regression problems (Polson et al. 2013). Nonetheless, even if this algorithm has been proved to be uniformly ergodic by Choi and Hobert (2013), the corresponding ergodicity constant depends exponentially on the number of observations  $n$  and on the dimension of the regression coefficients vector  $d$ .

To tackle these limitations, we propose to relax the constraint (2) and consider an *approximate* DA model. This will permit the choice of an augmented density with more

flexibility, fix the issues associated to the initial model and make inference more efficient in some cases. To this purpose, Section 2.2 presents the so-called AXDA framework which embeds approximate DA models controlled by a positive scalar parameter  $\rho$ . These models become asymptotically exact when  $\rho$  tends towards 0. Of course, some assumptions will be required on the approximate augmented density to guarantee a good approximation. The quality of this approximation will be assessed in Section 4 with non-asymptotic theoretical results.

## 2.2 Model

Instead of searching for an exact data augmentation scheme (2), some auxiliary variables  $\mathbf{z}$  can be introduced in order to define an approximate but asymptotically exact probability distribution. One possibility is to introduce an augmented distribution depending on a parameter  $\rho > 0$  and such that the associated marginal density defined by

$$\pi_\rho(\boldsymbol{\theta}) = \int_{\mathcal{Z}} \pi_\rho(\boldsymbol{\theta}, \mathbf{z}) d\mathbf{z}, \quad (3)$$

satisfies the following property.

**Property 1.** *For all  $\boldsymbol{\theta} \in \Theta$ ,  $\lim_{\rho \rightarrow 0} \pi_\rho(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ .*

By applying Scheffé's lemma (Scheffé 1947), this property yields the convergence in total variation, that is  $\|\pi_\rho - \pi\|_{\text{TV}} \rightarrow 0$  as  $\rho \rightarrow 0$ . A natural question is: how to choose the augmented density in (3) such that Property 1 is met? In this paper, we assume that  $\mathcal{Z} = \Theta$  and investigate AXDA schemes associated to an initial density (1) and defined by the approximate augmented density

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}) = \pi(\mathbf{z}) \kappa_\rho(\mathbf{z}, \boldsymbol{\theta}), \quad (4)$$

where  $\kappa_\rho$  is such that (4) defines a proper density.

**Remark 1.** *When  $\pi$  stands for a product of  $J$  densities, that is  $\pi \propto \prod_{j=1}^J \pi_j$ , the proposed approximate model can naturally be generalized to  $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:J}) \propto \prod_{j=1}^J \pi_j(\mathbf{z}_j) \kappa_\rho(\mathbf{z}_j, \boldsymbol{\theta})$ . Such a generalization will for instance be considered in Sections 3.1 and 3.2.*

The introduction of the proposed model (4) is aimed at avoiding a case-by-case search of an appropriate augmented approach. Although there might exist other marginal densities  $\pi_\rho$  satisfying Property 1, we restrict our analysis to models where  $\kappa_\rho(\cdot, \boldsymbol{\theta})$  weakly converges towards the Dirac measure at  $\boldsymbol{\theta}$  as  $\rho \rightarrow 0$  (Aguirregabiria et al. 2002). This is a sufficient condition to satisfy Property 1. In the sequel, we will call AXDA any approach based on (4) and satisfying these properties.

A natural choice for  $\kappa_\rho$  is to consider a standard kernel  $K$  (Wand and Jones 1995). Based on the latter, we define for all  $\mathbf{z}, \boldsymbol{\theta} \in \Theta$ ,  $\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto_{\mathbf{z}} \rho^{-d} K(\rho^{-1}(\boldsymbol{\theta} - \mathbf{z}))$  (Dang and Ehrhardt 2012). Beyond standard kernels but motivated by the same idea of measuring the discrepancy between  $\mathbf{z}$  and  $\boldsymbol{\theta}$ , one can also build on divergence functions widely used in the optimization literature to define a potentially asymmetric density  $\kappa_\rho$  such that for all  $\mathbf{z}, \boldsymbol{\theta} \in \Theta$ ,  $\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto_{\mathbf{z}} \exp(-\rho^{-1}\phi(\mathbf{z}, \boldsymbol{\theta}))$  where  $\phi$  is a strictly convex function w.r.t.  $\mathbf{z}$  admitting a unique minimizer  $\mathbf{z}^* = \boldsymbol{\theta}$  (Ben-Tal et al. 2001; Krichene et al. 2015; Fellows et al. 2019). Specific instances of such potentials are Bregman divergences such as the logistic loss and the Kullback-Leibler divergence, see Definition 1.

**Definition 1** (Bregman divergence). *Let  $\psi$  a continuously-differentiable and strictly convex function defined on a closed convex set. The Bregman divergence associated to  $\psi$  is defined by*

$$d_\psi(\mathbf{z}, \boldsymbol{\theta}) = \psi(\mathbf{z}) - \psi(\boldsymbol{\theta}) - \nabla\psi(\boldsymbol{\theta})^T(\mathbf{z} - \boldsymbol{\theta}). \quad (5)$$

Additional details associated to standard kernels and Bregman divergences are given in Section 2 in the supplementary material.

### 3 Benefits of AXDA by revisiting existing models

Before providing theoretical guarantees for AXDA models, this section proposes to review some important state-of-the-art works from the AXDA perspective described in Section 2. We do not pretend to give new insights about these approaches. We rather use them to illustrate potential benefits that can be gained by resorting to the proposed framework. For sake of clarity, these benefits are directly highlighted in the title of the following sections before being discussed in the latter.

### 3.1 Tractable posterior inference

This first section illustrates how an AXDA approach can alleviate the intractability of an initial posterior distribution  $\pi$  and significantly aid in the computations.

To this purpose, we consider the case where the posterior distribution  $\pi$  is intractable. Such a model for instance appears when  $\pi$  involves a constraint on some set (Liechty et al. 2009), admits a non-standard potential function such as the total variation norm (Chambolle et al. 2010; Pereyra 2016; Vono et al. 2019) or yields complicated conditional posterior distributions (Holmes and Mallick 2003). To simplify the inference, the aforementioned authors have considered special instances of AXDA by relying on an additional level involving latent variables  $\mathbf{z}$ , leading a hierarchical Bayesian model. In these cases, AXDA has been invoked in order to move a difficulty to the conditional posterior of  $\mathbf{z}$  where it can be dealt with more easily by using standard inference algorithms, see Section 5 in the supplementary material for more details. The following example, derived from Holmes and Mallick (2003), illustrates this idea.

**Example 1.** Let  $\mathbf{y} \in \mathbb{R}^n$  be a set of observations and  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times d}$  a design matrix filled with covariates. We consider a generalized non-linear model which writes

$$y_i | \boldsymbol{\theta} \sim p(y_i | g^{-1}(h(\mathbf{x}_i, \boldsymbol{\theta})), \sigma^2), \quad \forall i \in [n], \quad (6)$$

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta} | \mathbf{0}_d, \nu^2 \mathbf{I}_d), \quad (7)$$

where  $p$  belongs to the exponential family and has mean  $g^{-1}(h(\mathbf{x}_i, \boldsymbol{\theta}))$  and variance  $\sigma^2$  where  $g$  is a link function. As in classical regression problems, we are interested in inferring the regression coefficients  $\boldsymbol{\theta} \in \mathbb{R}^d$ . In the sequel, we set the non-parametric model  $h$  to be

$$h(\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^k \theta_j B(\mathbf{x}_i, \mathbf{k}_j), \quad (8)$$

where  $B(\mathbf{x}_i, \mathbf{k}_j)$  is a non-linear function of  $\mathbf{x}_i$  (e.g., regression splines) and  $\mathbf{k}_j$  is the knot location of the  $j$ -th basis. The difficulty here is the non-linearity of  $h$  which, combined with the non-Gaussian likelihood, rules out the use of efficient simulation schemes to sample from the posterior  $\pi(\boldsymbol{\theta} | \mathbf{y})$ . In order to mitigate this issue, Holmes and Mallick (2003) proposed to rely on an additional level which boils down to consider the approximate model



(4). More specifically, the aforementioned authors treated the non-linear predictor  $h$  as a Gaussian random latent variable which leads to the approximate model

$$y_i | z_i \sim p(y_i | g^{-1}(z_i), \sigma^2), \quad \forall i \in [n], \quad (9)$$

$$z_i | \boldsymbol{\theta} \sim \mathcal{N}(z_i | h(\mathbf{x}_i, \boldsymbol{\theta}), \rho^2), \quad \forall i \in [n], \quad (10)$$

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta} | \mathbf{0}_d, \nu^2 \mathbf{I}_d). \quad (11)$$

Here, AXDA has been applied only to the likelihood function with  $\kappa_\rho$  chosen as the univariate normal distribution (10) leading to a smoothed likelihood function. The main advantage of relying on such a model is that the posterior conditional distribution  $\pi_\rho(\boldsymbol{\theta} | \mathbf{z}, \mathbf{X})$ , with  $\mathbf{z} = [z_1, \dots, z_n]^T$ , is now a multivariate normal distribution. In addition, by moving the difficulty induced by  $h$  to the conditional posterior of  $z_i$ , we are now dealing with a generalized linear model where standard techniques can be applied (Albert and Chib 1993; Polson et al. 2013).

Beyond the widely-used Gaussian choice for  $\kappa_\rho$  (Holmes and Mallick 2003; Liechty et al. 2009; Barbos et al. 2017; Vono et al. 2019), more general AXDA approaches can be built by taking inspiration from these works. To this purpose, we recommend to choose  $\kappa_\rho$  w.r.t. the prior and likelihood at stake. For instance, when a Poisson likelihood function and a complex prior distribution on its intensity  $\boldsymbol{\theta}$  are considered, one option for  $\phi$  (see Section 2.2) would be an Itakura-Saito divergence since it preserves the positivity constraint on  $\boldsymbol{\theta}$  and yields the well-known Gamma-Poisson model (Canny 2004).

## 3.2 Distributed inference

When data are stored on multiple machines and/or one is interested in respecting their privacy, this section illustrates how AXDA can be resorted to perform distributed computations.

Let consider observed data  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , where  $\mathbf{x}_i$  stands for the covariates associated to observation  $y_i$ , which are distributed among  $B$  nodes within a cluster. By adopting a prior  $\nu(\boldsymbol{\theta})$  and by assuming that the likelihood can be factorized w.r.t. the  $B$  nodes, the posterior distribution of the variable of interest  $\boldsymbol{\theta}$  writes

$$\pi(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \propto \nu(\boldsymbol{\theta}) \prod_{b=1}^B \prod_{i \in \text{node } b} \exp(-f_i(y_i; h(\mathbf{x}_i, \boldsymbol{\theta}))). \quad (12)$$

Such models classically appear in statistical machine learning when generalized linear models (GLMs) (Dobson and Barnett 2008) are considered. In these cases,  $h(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{x}_i^T \boldsymbol{\theta}$ . Due to the distributed environment, sampling efficiently from (12) is challenging and a lot of “divide-and-conquer” approaches have been proposed in the past few years to cope with this issue (Wang and Dunson 2013; Scott et al. 2016). These methods launch independent Markov chains on each node  $b$  and then combine the outputs of these local chains to obtain an approximation of the posterior of interest (12). Nonetheless, the averaging schemes used to combine the local chains might lead to poor approximations when  $\pi$  is high-dimensional and non-Gaussian. Instead, considering a special instance of AXDA circumvents the previously mentioned drawbacks by introducing local auxiliary variables on each node such that

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z} | \mathbf{y}, \mathbf{X}) \propto \nu(\boldsymbol{\theta}) \prod_{b=1}^B \prod_{i \in \text{node } b} \exp(-f_i(y_i; z_i)) \kappa_\rho(z_i, h(\mathbf{x}_i, \boldsymbol{\theta})). \quad (13)$$

The posterior distribution of the auxiliary variables conditionally to  $\boldsymbol{\theta}$  only depends on the data available at a given node. Based on this nice property, the joint posterior can be sampled efficiently with a Gibbs sampler, see Rendell et al. (2020) for a comprehensive review. We emphasize that the benefits described in this section for Monte Carlo sampling also hold when one wants to use other types of algorithms (e.g., expectation-maximization or variational Bayes), see Section 5 in the supplementary material.

### 3.3 Robust inference

By noting that classical robust hierarchical models fall into the proposed framework, this section shows that AXDA is also a relevant strategy to perform robust inference by coping with model misspecification by modeling additional sources of uncertainty.

Considering a well-chosen *demarginalization* procedure is known to yield robustness properties in some cases (Robert and Casella 2004). Some approaches took advantage of this idea in order to build robust hierarchical Bayesian models w.r.t. possible outliers in the data. For instance, such models can be built by allowing each observation to be randomly drawn from a local statistical model, as described in the recent review of Wang and Blei (2018). This “localization” idea is illustrated in Figure 1. Many of these models

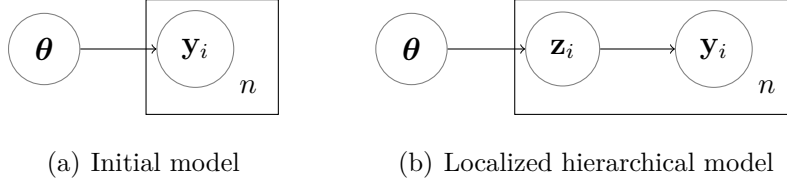


Figure 1: Concept of localization. Comparison between the initial (left) and the localized hierarchical Bayesian (right) models with  $n$  the number of observations  $y_i$ .

can be viewed as particular instances of AXDA. Indeed, assume that  $n$  data points  $y_i$  are independently and identically distributed (i.i.d.) defining the likelihood function

$$\pi(\mathbf{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^n \pi(y_i|\boldsymbol{\theta}), \quad (14)$$

where  $\boldsymbol{\theta} \in \Theta$  is a common parameter. Applying AXDA as described in Section 2 by introducing  $n$   $d$ -dimensional auxiliary variables stacked into the vector  $\mathbf{z}_{1:n}$  leads to the augmented likelihood

$$\pi_\rho(\mathbf{y}, \mathbf{z}_{1:n}|\boldsymbol{\theta}) \propto \prod_{i=1}^n \pi(y_i|\mathbf{z}_i)\kappa_\rho(\mathbf{z}_i, \boldsymbol{\theta}). \quad (15)$$

The statistical model defined by (15) implies a hierarchical Bayesian model similar to the localized one depicted on Figure 1(b) and corresponds in general to an approximation of the initial one, see Example 2.

**Example 2. Robust logistic regression** – Assume that for all  $i \in [n]$ ,  $\pi(y_i|\boldsymbol{\theta}) = \mathcal{B}\left(\sigma(\mathbf{x}_i^T \boldsymbol{\theta})\right)$ , where  $\mathcal{B}$  stands for the Bernoulli distribution,  $\sigma$  for the sigmoid function,  $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$  for the transpose of the design matrix and  $\boldsymbol{\theta}$  for the regression coefficients vector to infer. Then as proposed by Wang and Blei (2018), one can robustify the inference by assuming that each observation  $y_i$  is drawn from a local and independent model  $\mathcal{B}\left(\sigma(\mathbf{x}_i^T \mathbf{z}_i)\right)$  associated to an auxiliary parameter  $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\theta}, \rho^2 \mathbf{I}_d)$ . In this case,  $\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto \prod_{i=1}^n \mathcal{N}(\mathbf{z}_i | \boldsymbol{\theta}, \rho^2 \mathbf{I}_d)$ .

Beyond the convenient Gaussian prior  $\kappa_\rho$ , its choice can be motivated by robust loss functions. In the statistical machine learning literature, the absolute or Huber losses are of common use (She and Owen 2011). In Bayesian linear inverse problems considered in the signal processing community, it is classical to approximate a complicated forward physical

model in order to yield tractable computations. If the latter can be written as  $\mathbf{y} = h(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$ , with  $\boldsymbol{\epsilon} \sim \pi(\boldsymbol{\epsilon})$ , then introducing a latent variable  $\mathbf{z} \sim \kappa_\rho(\mathbf{z}, h(\boldsymbol{\theta}))$  such that  $\mathbf{y} = \mathbf{z} + \boldsymbol{\epsilon}$  allows to take into consideration the model approximation. In those cases, one can set  $\kappa_\rho$  to be the distribution of the modeling error which could be adjusted thanks to some expertise.

### 3.4 Inheriting sophisticated inference schemes from ABC

Finally, this section shows that AXDA models, by sharing strong connections with ABC, might inherit sophisticated algorithms to sample from (4).

ABC stands for a family of methods that permit to cope with intractable likelihoods by sampling from the latter instead of evaluating them. In a nutshell, if one’s goal is to infer a parameter  $\boldsymbol{\theta}$  based on a posterior of interest, the simplest ABC rejection sampler is as follows. At iteration  $t$ , draw a candidate  $\boldsymbol{\theta}^{(t)}$  from the prior, generate pseudo-observations  $\mathbf{z}$  from the likelihood given this candidate and accept  $\boldsymbol{\theta}^{(t)}$  if  $\mathbf{z} = \mathbf{y}$  where  $\mathbf{y}$  is the observations vector. Many more sophisticated ABC samplers have been derived. We refer the interested reader to the recent review by [Sisson et al. \(2018a\)](#) for more information about ABC methods.

Among a huge literature on ABC (also called likelihood-free) methods, *noisy ABC* approaches proposed and motivated by [Fearnhead and Prangle \(2012\)](#) and [Wilkinson \(2013\)](#) are strongly related to AXDA. Indeed, only comparing the underlying models, AXDA with observation splitting is equivalent to noisy ABC. To see this, let  $\pi(\mathbf{y}|\boldsymbol{\theta})$  stand for an intractable likelihood. Noisy ABC replaces the exact inference based on  $\pi$  by considering the pseudo-likelihood with density

$$\pi_\rho(\mathbf{y}|\boldsymbol{\theta}) \triangleq \int_{\Theta} \pi_\rho(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int_{\Theta} \pi(\mathbf{z}|\boldsymbol{\theta}) \kappa_\rho(\mathbf{z}, \mathbf{y}) d\mathbf{z}. \quad (16)$$

This density has exactly the same formulation as the one defined in (4) except that noisy ABC splits the observations  $\mathbf{y}$  instead of the parameter of interest  $\boldsymbol{\theta}$ . Capitalizing on this equivalence property, also pointed out by [Rendell et al. \(2020\)](#), one can derive efficient algorithms for AXDA from the ABC framework. For instance, [Rendell et al. \(2020\)](#) recently built on the works of [Beaumont et al. \(2002\)](#); [Del Moral et al. \(2012\)](#) in the ABC context to propose a bias correction approach and a sequential Monte Carlo (SMC) algorithm avoiding

the tuning of the tolerance parameter  $\rho$ . Obviously, many other inspirations from ABC can be considered, such as the parallel tempering approach of [Baragatti et al. \(2013\)](#) among others, to make the inference from an AXDA model more flexible and efficient.

## 4 Theoretical guarantees

By building on existing approaches, Section 3 showed that AXDA can be used in quite general and different settings depending on ones motivations. In order to further promote the use of such approximate augmented models, this section goes beyond the empirical bias analysis performed by previous works and provides quantitative bounds on the error between the initial and the approximate model. More precisely, for a fixed tolerance parameter  $\rho > 0$ , non-asymptotic results on the error associated to densities, potentials and credibility regions are derived. We will assume all along this section that  $\Theta = \mathbb{R}^d$ . The proofs of the results of this section can be found in Section 1 of the supplementary material.

### 4.1 Results for standard kernels

In this section, we consider the case  $\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto \rho^{-d} K(\rho^{-1}(\boldsymbol{\theta} - \mathbf{z}))$  where  $K$  is a kernel, see Section 2.2. Under this model,  $\pi_\rho$  stands for the convolution of  $\pi$  and  $\kappa_\rho$  and the following results hold.

**Proposition 1.** *Let  $\pi \in L^1$ . The marginal with density  $\pi_\rho$  in (3) has the following properties.*

*i) Let  $\pi$  stand for a pdf associated to the random variable  $\boldsymbol{\theta}$  and  $\mathbb{E}_{\kappa_\rho}(X) = 0$ . Then, the expectation and variance under  $\pi_\rho$  are given by*

$$\mathbb{E}_{\pi_\rho}(\theta) = \mathbb{E}_\pi(\theta) \tag{17}$$

$$\text{var}_{\pi_\rho}(\theta) = \text{var}_\pi(\theta) + \text{var}_{\kappa_\rho}(\theta). \tag{18}$$

*ii)  $\text{supp}(\pi_\rho) \subseteq S$  where  $S$  is the closure of  $\{\mathbf{x} + \mathbf{z}; \mathbf{x} \in \text{supp}(\pi), \mathbf{z} \in \text{supp}(\kappa_\rho)\}$ . The notation  $\text{supp}(h) = \{\mathbf{x} \in \mathcal{X} \mid h(\mathbf{x}) \neq 0\}$  refers to the support of a function  $h : \mathcal{X} \rightarrow \mathbb{R}$ .*

*iii) If both  $\pi$  and  $\kappa_\rho$  are log-concave, then  $\pi_\rho$  is log-concave.*

*iv) If  $\kappa_\rho \in \mathcal{C}^\infty(\mathbb{R}^d)$  and  $|\partial^k \kappa_\rho|$  is bounded for all  $k \geq 0$ , then  $\pi_\rho$  is infinitely differentiable w.r.t.  $\boldsymbol{\theta}$ .*

Proposition 1 permits to draw several conclusions about the inference based on  $\pi_\rho$ . Firstly, the infinite differentiability of  $\pi_\rho$  (Property *iv*) implies that it stands for a smooth approximation of  $\pi$ , see Figure 5 in Section 5.2. Secondly, Property *i*) of Proposition 1 is reassuring regarding the inference task. Indeed, if  $\pi$  stands for a prior distribution, then considering the approximation  $\pi_\rho$  simply corresponds to a more diffuse prior knowledge around the same expected value, see Section 5.2. Thus, more weight will be given to the likelihood if a posterior distribution is derived with this prior. On the other hand, if  $\pi$  stands for a likelihood, then considering the approximation  $\pi_\rho$  yields the opposite behavior: the likelihood becomes less informative w.r.t. the prior. This idea is directly related to robust hierarchical Bayesian models discussed in Section 3.3.

We now provide quantitative bounds on the approximation implied by considering the marginal  $\pi_\rho$  instead of  $\pi$ . For  $p \geq 1$ , we define the  $p$ -Wasserstein distance between  $\pi$  and  $\pi_\rho$  by

$$W_p(\pi, \pi_\rho) = \left( \min_{\mu} \left\{ \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|_2^p d\mu(\mathbf{z}, \boldsymbol{\theta}); \mu \in \Gamma(\pi_\rho, \pi) \right\} \right)^{1/p}, \quad (19)$$

where  $\Gamma(\pi_\rho, \pi)$  is the set of probability distributions  $\mu(\boldsymbol{\theta}, \mathbf{z})$  with marginals  $\pi_\rho$  and  $\pi$  w.r.t.  $\boldsymbol{\theta}$  and  $\mathbf{z}$ , respectively. Under mild assumptions on the kernel  $K$ , Proposition 2 gives a simple and practical upper bound on (19).

**Proposition 2.** *Assume that  $\pi_\rho$  in (3) stands for a pdf associated to the variable  $\boldsymbol{\theta}$ . Let  $p \geq 1$  such that  $m_p \triangleq \left( \int_{\mathbb{R}^d} \|\mathbf{u}\|_2^p K(\mathbf{u}) d\mathbf{u} \right)^{1/p} < \infty$ . Then, we have*

$$W_p(\pi, \pi_\rho) \leq \rho m_p. \quad (20)$$

Note that (20) holds without assuming additional assumptions on the initial density  $\pi$  such as infinite differentiability. If the latter is assumed w.r.t. the parameter of interest  $\boldsymbol{\theta}$ , then one can estimate the bias  $\pi - \pi_\rho$  with a Taylor expansion of  $\pi$  similarly to bias analysis in ABC, see Sisson et al. (2018b). Table 1 gives closed-form expressions of  $m_2$  for the multivariate generalizations of standard kernels. One can see that the constant  $m_2$

Table 1: Closed-form expressions of  $m_2$  appearing in (20) for multivariate generalizations of standard kernels where  $d$  denotes the dimension.

	Gaussian	Cauchy	Laplace	Dirichlet	Uniform	Triangular	Epanechnikov
$m_2$	$\sqrt{d}$	-	$\sqrt{2d}$	-	$\sqrt{d/3}$	$\sqrt{d/6}$	$\sqrt{d/5}$

has the same dependence w.r.t. the dimension  $d$  for the considered standard kernels  $K$ . Hence, in high-dimensional scenarios, the approximation quality will be more affected by an inappropriate value for the tolerance parameter  $\rho$  rather than by the choice of  $K$ . In Section 5, we illustrate Proposition 2 with numerical experiments.

## 4.2 Pointwise bias for Bregman divergences

In complement to Section 4.1 where  $\kappa_\rho$  was built using kernels, we now analyze the bias induced by considering  $\pi_\rho$  when  $\kappa_\rho$  is derived from a Bregman divergence  $d_\psi$  (see Definition 1), that is

$$\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto \exp\left(-\frac{d_\psi(\mathbf{z}, \boldsymbol{\theta})}{\rho}\right). \quad (21)$$

Under regularity assumptions on both  $\pi$  and  $\kappa_\rho$ , Proposition 3 shows the dependence of the pointwise bias  $\pi_\rho - \pi$  w.r.t. to the tolerance parameter  $\rho$  when the latter is sufficiently small.

**Proposition 3.** *Assume that  $\pi$  is analytic and twice differentiable on  $\mathbb{R}^d$  and so does  $d_\psi$  w.r.t. its first argument. Let  $\boldsymbol{\theta} \in \mathbb{R}^d$  such that both  $\mathbf{H}_\pi(\boldsymbol{\theta})$  and  $\mathbf{H}_{d_\psi}(\boldsymbol{\theta})^{-1}$  exist and are continuous, where  $\mathbf{H}_\pi(\boldsymbol{\theta})$  is the Hessian matrix of  $\pi$  and  $\mathbf{H}_{d_\psi}(\boldsymbol{\theta}) \triangleq \frac{\partial^2 d_\psi(\mathbf{z}, \boldsymbol{\theta})}{\partial \mathbf{z}^2} \Big|_{\mathbf{z}=\boldsymbol{\theta}}$  is the Hessian matrix associated to  $d_\psi(\cdot, \boldsymbol{\theta})$ . Then, if*

- $\|\mathbf{H}_\pi\| \leq C < \infty$
- $\|\mathbf{H}_{d_\psi}\| \geq c > 0$ ,

it follows that

$$\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{\rho}). \quad (22)$$

In addition, if we have  $\int_{\mathbb{R}^d} \mathbf{u} \kappa_\rho(\boldsymbol{\theta} - \sqrt{\rho} \mathbf{u}, \boldsymbol{\theta}) d\mathbf{u} = \mathbf{0}_d$ , then

$$\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}) = \frac{\rho}{2} \text{Trace} \left( \mathbf{H}_\pi(\boldsymbol{\theta}) \mathbf{H}_{d_\psi}(\boldsymbol{\theta})^{-1} \right) + o(\rho). \quad (23)$$

Note that when  $\psi(\mathbf{z}) = \|\mathbf{z}\|_2^2/2$ ,  $\kappa_\rho$  stands for a Gaussian smoothing kernel, see Section 4.1. In that case, we have the sanity check that the dependence w.r.t.  $\rho$  of the bias between  $\pi$  and  $\pi_\rho$  in (23) is the same as the one derived by [Sisson et al. \(2018b\)](#) when interpreting  $\kappa_\rho$  as a kernel.

### 4.3 A detailed non-asymptotic analysis for Gaussian smoothing

The previous sections gave quantitative approximation results for a large class of densities  $\kappa_\rho$  built either via a kernel or a Bregman divergence. In this section, we provide complementary results by restricting our analysis on the case

$$\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{z} | \boldsymbol{\theta}, \rho^2 \mathbf{I}_d). \quad (24)$$

This particular yet convenient assumption will allow to complement and sharpen results of Section 4.1 by deriving quantitative bounds which take into account the regularity properties of  $f$ . Furthermore, these bounds can be extended to a sum of potential functions  $f = \sum_i f_i$  and used to assess the bias associated to both log-densities and credibility regions. This analysis is also motivated by the fact that the Gaussian smoothing case has been widely advocated in the literature since it generally leads to simple inference steps ([Holmes and Mallick 2003](#); [Giovannelli 2008](#); [Liechty et al. 2009](#); [Dümbgen and Rufibach 2009](#)), and can be related to both the ADMM in optimization ([Boyd et al. 2011](#); [Vono et al. 2019](#)) and the approximation involved in proximal MCMC methods ([Pereyra 2016](#); [Durmus et al. 2018](#); [Salim et al. 2019](#)). Unfortunately, a straightforward generalization of the proof techniques used in the sequel does not give informative upper bounds for smoothing associated to other Bregman divergences.

#### 4.3.1 Assumptions

To derive non-asymptotic bounds between quantities related to  $\pi_\rho$  defined in (3) and  $\pi$  in (1), some complementary assumptions on  $f = -\log \pi$  will be required. They are detailed



hereafter. For simplicity and with a little abuse of notations, we also denote here by  $f(\boldsymbol{\theta})$  the potential associated to (1) when  $\pi(\mathbf{y}|\boldsymbol{\theta})$  stands for a likelihood.

(A<sub>1</sub>)  $f$  is  $L_f$ -Lipschitz w.r.t.  $\|\cdot\|_2$ , that is  $\exists L_f \geq 0$  such that for all  $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^d$ ,  $|f(\boldsymbol{\theta}) - f(\boldsymbol{\eta})| \leq L_f \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_2$ . When  $\pi$  is a likelihood, it is further assumed that  $L_f$  is independent of  $\mathbf{y}$ .

(A<sub>2</sub>)  $f$  is continuously differentiable and has an  $M_f$ -Lipschitz continuous gradient w.r.t.  $\|\cdot\|_2$ , that is  $\exists M_f \geq 0$  such that for all  $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^d$ ,  $\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\eta})\|_2 \leq M_f \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_2$ .

(A<sub>3</sub>)  $f$  is convex, that is for every  $\alpha \in [0, 1]$ ,  $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^d$ ,  $f(\alpha\boldsymbol{\theta} + (1 - \alpha)\boldsymbol{\eta}) \leq \alpha f(\boldsymbol{\theta}) + (1 - \alpha)f(\boldsymbol{\eta})$ .

(A<sub>4</sub>)  $M_f = \int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{\theta})\|_2^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty$ .

Assumptions (A<sub>1</sub>), (A<sub>2</sub>) and (A<sub>3</sub>) on the potential  $f$  stand for standard regularity assumptions in the optimization literature and cover a large class of functions  $f$  (Beck and Teboulle 2009; Bolte et al. 2014). In the broad statistical community, (A<sub>1</sub>) has been used by Durmus et al. (2018) to derive non-asymptotic bounds on the total variation distance between probability distributions while (A<sub>2</sub>) stands for a sufficient condition to have a strong solution to the overdamped Langevin stochastic differential equation (Durmus and Moulines 2017).

Under the previous assumptions (not used all at once), non-asymptotic upper bounds on the total variation distance between  $\pi_\rho$  and  $\pi$  are derived in Section 4.3.2. Then, Sections 4.3.3 and 4.3.4 take advantage of this bound to state theoretical properties on the potential functions and credibility regions.

### 4.3.2 Non-asymptotic bounds on the total variation distance

In this section, we make additional regularity assumptions on the potential  $f$  in order to show quantitative results depending explicitly on regularity constants associated to  $f$ . Two different cases will be considered, namely Lipschitz potentials, and differentiable, gradient-Lipschitz and convex ones.

**Lipschitz potential** – When the potential function  $f$  is assumed to be Lipschitz continuous but not necessarily continuously differentiable, the following result holds.

**Theorem 1.** *Let a potential function  $f$  satisfy  $(A_1)$ . Then,*

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq 1 - \Delta_d(\rho), \quad (25)$$

where

$$\Delta_d(\rho) = \frac{D_{-d}(L_f \rho)}{D_{-d}(-L_f \rho)}. \quad (26)$$

The function  $D_{-d}$  is a parabolic cylinder function defined for all  $d > 0$  and  $z \in \mathbb{R}$  by

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz-x^2/2} x^{d-1} dx. \quad (27)$$

As expected from Property 1, note that this bound tends towards zero when  $\rho \rightarrow 0$ . Additionally, this bound depends on few quantities that can be computed, bounded or approximated in real applications: the dimension of the problem  $d$ , the Lipschitz constant  $L_f$  associated to the regularized potential  $f$  and the tolerance parameter  $\rho$ . In the limiting case  $\rho \rightarrow 0$ , the following equivalent function for the upper bound derived in (25) holds.

**Corollary 1.** *In the limiting case  $\rho \rightarrow 0$ , we have:*

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq \rho L_f \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} + o(\rho), \quad (28)$$

where for all  $z > 0$  as  $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx$ .

Under some regularity conditions (here Lipschitz continuity) on the potential function  $f$ , Proposition 1 states that  $\|\pi_\rho - \pi\|_{\text{TV}}$  grows at most linearly w.r.t. the parameter  $\rho$  and w.r.t.  $L_f$  when  $\rho$  is sufficiently small. Moreover, using Stirling-like approximations when  $d$  is large in the equivalence relation (28) may give a mild dependence on the dimensionality of the problem in  $\mathcal{O}(L_f d^{1/2})$ . Potential functions verifying the hypothesis of Theorem 1 are common in machine learning and signal/image processing problems, see Section 3 in the online supplementary material. As an archetypal example, the sparsity promoting potential function defined for all  $\boldsymbol{\theta} \in \mathbb{R}^d$  by  $f(\boldsymbol{\theta}) = \tau \|\boldsymbol{\theta}\|_1$  with  $\tau > 0$  is Lipschitz continuous with Lipschitz constant  $L_f = \tau\sqrt{d}$  and satisfies Theorem 1 and Proposition 1. In this case,

the dependence of (28) is linear w.r.t.  $d$  when  $d$  is large and  $\rho$  is small. Note also that continuously differentiable functions on a compact set are Lipschitz continuous.

**Convex and gradient-Lipschitz potential** – We now show a complementary result by assuming  $f$  to be convex and continuously differentiable with a Lipschitz-continuous gradient.

**Theorem 2.** *Let a potential function  $f$  satisfy  $(A_2)$ ,  $(A_3)$  and  $(A_4)$ . Then, when  $\pi$  stands for a pdf associated to  $\boldsymbol{\theta}$ , we have:*

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq 1 - \frac{1}{(1 + 2\rho^2 M_f)^{d/2}} \left( 1 - \frac{\rho^4 M_f M_f}{1 + 2\rho^2 M_f} \right). \quad (29)$$

In the limiting case  $\rho \rightarrow 0$ , the upper bound in (29) has a simpler expression as shown hereafter.

**Corollary 2.** *In the limiting case  $\rho \rightarrow 0$ , we have:*

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq \rho^2 d M_f + o(\rho^2). \quad (30)$$

Note that the dependences w.r.t. both  $\rho$  and  $d$  in Corollary 1 and 2 are similar to the ones found by [Nesterov and Spokoiny \(2017\)](#) for optimization purposes.

Figure 2 gives the behavior of the upper bounds in (25) and (29) w.r.t. the dimensionality  $d$  of the problem ranging from 1 to  $10^6$  and as a function of  $\rho$  in log-log scale. The linear (resp. quadratic) relation between this upper bound and  $\rho$  shown in (28) (resp. (30)) is clearly observed for small values of  $\rho$ . Nonetheless, these upper bounds are not a silver bullet. Indeed, as expected, for a fixed value of the parameter  $\rho$ , the approximation error increases as the dimension  $d$  grows. Thus, these bounds suffer from the curse of dimensionality and become non-informative in high-dimension if  $\rho$  is not sufficiently small.

Theorem 1 is easily extended to the case where the initial density  $\pi$  is expressed as a product of several terms. If  $\pi$  stands for the pdf associated to the variable  $\boldsymbol{\theta}$ , this boils down to considering

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^J \pi_j(\boldsymbol{\theta}) \propto \exp \left( - \sum_{j=1}^J f_j(\boldsymbol{\theta}) \right), \quad (31)$$

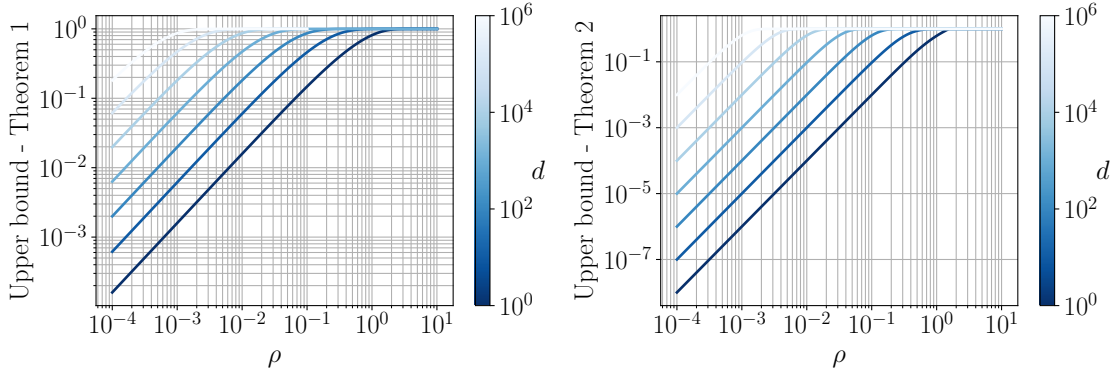


Figure 2: Behavior of the quantitative bounds shown in Theorems 1 and 2 w.r.t.  $\rho$  in log-log scale for a set of dimensions  $d$ . The other quantities appearing in the bounds have been set to 1.

where for all  $j \in [J]$ ,  $f_j : \mathbb{R}^d \rightarrow \mathbb{R}$ , and a natural generalization of AXDA when applied to each  $\pi_j$ , which writes

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:J}) \propto \prod_{j=1}^J \pi_j(\mathbf{z}_j) \kappa_{\rho_j}(\mathbf{z}_j; \boldsymbol{\theta}) \propto \exp\left(-\sum_{j=1}^J f_j(\mathbf{z}_j) + \frac{1}{2\rho_j^2} \|\mathbf{z}_j - \boldsymbol{\theta}\|_2^2\right). \quad (32)$$

Under this product form, we have the following corollary.

**Corollary 3.** *For all  $j \in [J]$ , let  $f_j$  satisfy  $(A_1)$ . Then,*

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq 1 - \prod_{j=1}^J \Delta_d^{(j)}(\rho_j), \quad (33)$$

where  $\Delta_d^{(j)}(\rho_j) = D_{-d}(L_{f_j}\rho_j)/D_{-d}(-L_{f_j}\rho_j)$ .

Unfortunately, Theorem 2 cannot be extended to the multiple splitting scenario. We are nevertheless confident that quantitative bounds can be found with different proof techniques but this task goes beyond the scope of this paper.

### 4.3.3 Uniform bounds on potentials

From an optimization point of view, it is quite common to consider potential functions associated to densities. For such applications, we give hereafter a quantitative uniform bound on the difference between the potential functions associated to  $\pi$  and  $\pi_\rho$ . Similarly

to the definition of the potential function  $f$  in (1), we define the potential function  $f_\rho$  associated to the approximate marginal  $\pi_\rho$  in (3), for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ , by

$$f_\rho(\boldsymbol{\theta}) = -\log \int_{\mathbb{R}^d} \exp(-f(\mathbf{z})) \kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}. \quad (34)$$

By considering a Gaussian smoothing kernel  $\kappa_\rho$ , the potential  $f_\rho$  becomes

$$f_\rho(\boldsymbol{\theta}) = -\log \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2} \|\mathbf{z} - \boldsymbol{\theta}\|_2^2\right) d\mathbf{z} + \frac{d}{2} \log(2\pi\rho^2). \quad (35)$$

Note that  $f_\rho(\boldsymbol{\theta})$  appears as a regularized version of  $f(\boldsymbol{\theta})$ .

**Proposition 4.** *Let  $f$  satisfy  $(A_1)$ . Then, for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ ,*

$$L_\rho \leq f_\rho(\boldsymbol{\theta}) - f(\boldsymbol{\theta}) \leq U_\rho, \quad (36)$$

with

$$L_\rho = \log N_\rho - \log D_{-d}(-L_f\rho), \quad (37)$$

$$U_\rho = \log N_\rho - \log D_{-d}(L_f\rho), \quad (38)$$

and

$$N_\rho = \frac{2^{d/2-1} \Gamma(d/2)}{\Gamma(d) \exp(L_f^2 \rho^2 / 4)}. \quad (39)$$

It is easily observed that these bounds are informative in the limiting case  $\rho \rightarrow 0$  since they both tend towards zero.

#### 4.3.4 Uniform bounds on credibility regions

When  $\pi$  stands for the density associated to a posterior distribution, one advantage of Bayesian analysis is its ability to derive the underlying probability distribution of the variable of interest  $\boldsymbol{\theta}$  and thereby to provide credibility information under this distribution. This uncertainty information is particularly relevant and essential for real-world applications. Since the marginal  $\pi_\rho$  stands for an approximation of the original target distribution  $\pi$ , it is important to control the credibility regions under  $\pi_\rho$  w.r.t. those drawn under  $\pi$ . The control in total variation distance given by Theorem 1 is already a good indication. However, it is possible to quantify more precisely the difference between the credible regions (Robert 2001) with confidence level  $(1 - \alpha)$  under  $\pi_\rho$  and  $\pi$ , as stated below.

**Proposition 5.** *Let  $\pi$  be a posterior distribution associated to  $\boldsymbol{\theta}$  and  $f$  such that  $(A_1)$  is verified. Let  $\mathcal{C}_\alpha^\rho$  an arbitrary  $(1 - \alpha)$ -credibility region under  $\pi_\rho$ , that is  $\mathbb{P}_{\pi_\rho}(\boldsymbol{\theta} \in \mathcal{C}_\alpha^\rho) = 1 - \alpha$  with  $\alpha \in (0, 1)$ . Then,*

$$(1 - \alpha) \frac{N_\rho}{D_{-d}(-L_f \rho)} \leq \int_{\mathcal{C}_\alpha^\rho} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \min \left( 1, (1 - \alpha) \frac{N_\rho}{D_{-d}(L_f \rho)} \right), \quad (40)$$

where  $N_\rho$  is defined in (39).

Proposition 5 states that the coverage of  $\pi$  under  $\mathcal{C}_\alpha^\rho$  can be determined for a fixed value of  $\rho$ . Thus, it is even possible to obtain a theoretical comprehensive description of  $\mathcal{C}_\alpha^\rho$  w.r.t. the initial target density  $\pi$  before conducting an AXDA-based inference. The bounds in (40) permit to choose a parameter  $\rho$  in order to ensure a prescribed coverage property. The behavior of these bounds w.r.t.  $\rho$  is the same as in Section 4.3.2, i.e., linear behavior w.r.t.  $\rho$  when this parameter is sufficiently small.

## 5 Numerical illustrations

This section illustrates the quantitative results shown in Sections 4.1 and 4.3 on three different examples which classically appear in statistical signal processing and machine learning. As shown in Table 1, the bias induced by considering  $\pi_\rho$  is mostly driven by the value of the tolerance parameter  $\rho$  rather than by the choice of  $\kappa_\rho$ . Hence, for simplicity, most of the numerical illustrations hereafter consider the case where  $\kappa_\rho$  is a Gaussian smoothing kernel. Additional illustrations can be found in the online supplementary material.

### 5.1 Multivariate Gaussian example

We start by performing a sanity check with the simple case where  $\pi$  stands for a multivariate Gaussian density that is

$$\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (41)$$

where  $\boldsymbol{\Sigma}$  is assumed to be positive definite. If  $\kappa_\rho(\cdot, \boldsymbol{\theta})$  is taken to be Gaussian density with mean  $\boldsymbol{\theta}$  and covariance matrix  $\rho^2 \mathbf{I}_d$ , then one can show that

$$\pi_\rho(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\Sigma} + \rho^2 \mathbf{I}_d). \quad (42)$$

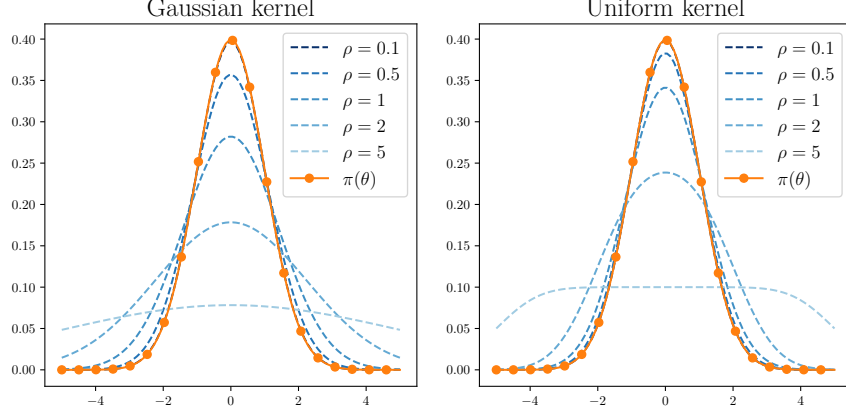


Figure 3: Bias between  $\pi_\rho$  and  $\pi$  in the case  $\Theta = \mathbb{R}$ ,  $\pi = \mathcal{N}(\mu, \sigma^2)$  with  $\mu = 0$  and  $\sigma = 1$ . (left)  $\pi_\rho$  is built with a Gaussian kernel  $\mathcal{N}(0, \rho^2)$  and (right) with a uniform kernel on  $[-\rho, \rho]$ . Note that the curves associated to  $\pi$  and  $\pi_\rho$  for  $\rho = 0.1$  are overlapping.

In particular, let consider the univariate setting, that is  $\Theta = \mathbb{R}$ ,  $\Sigma = \sigma^2$ . In this case, the variance under  $\pi_\rho$  is  $\sigma^2 + \rho^2$  and simply corresponds to the variance under  $\pi$  inflated by a factor  $\rho^2$ . Therefore, the approximation will be reasonable if  $\rho^2/\sigma^2$  is sufficiently small, see Figure 3. In this Figure, we also show the approximation induced by considering a uniform kernel instead of a Gaussian one. The smoothing via the uniform kernel performs slightly better than Gaussian smoothing due to its lower variance. In both cases, the approximation is reasonable for small  $\rho$  although  $\pi_\rho$ , built with a uniform kernel, no longer belongs to the Gaussian family.

In order to illustrate the proposed upper bounds on both 2-Wasserstein and total variation distances, we consider a covariance matrix  $\Sigma$  which stands for a squared exponential matrix commonly used in applications involving Gaussian processes (Higdon 2007) and which writes

$$\Sigma_{ij} = 2 \exp\left(-\frac{(s_i - s_j)^2}{2a^2}\right) + 10^{-6}\delta_{ij}, \forall i, j \in [d] \quad (43)$$

where  $a = 1.5$ ,  $s_{i,i \in [d]}$  are regularly spaced scalars on  $[-3, 3]$  and  $\delta_{ij} = 1$  if  $i = j$  and zero otherwise.

Figure 4 shows the behavior of the quantitative bounds derived in Proposition 2 and Theorem 2 for  $d \in \{10, 100\}$ . The Gaussian case allows to compute exactly  $W_2(\pi, \pi_\rho)$  by noting that  $W_2^2(\pi, \pi_\rho) = \text{Trace}(\Sigma + \rho^2 \mathbf{I}_d - 2\rho \Sigma^{1/2})$ . On the other hand,  $\|\pi - \pi_\rho\|_{\text{TV}}$  has been

estimated by using a Monte Carlo approximation. One can note that the general upper bound on the 2-Wasserstein distance is quite conservative for small  $\rho$  since it does not catch the behavior in  $\mathcal{O}(\rho^2)$  when  $\rho$  is small. This is essentially due to the fact that this bound only assumes a finite moment property and does not require any regularity assumptions on  $\pi$  such as differentiability or strong convexity of its potential. On the contrary, the bound on the total variation distance, derived under stronger assumptions, manages to achieve a rate of the order  $\mathcal{O}(\rho^2)$  for small  $\rho$ .

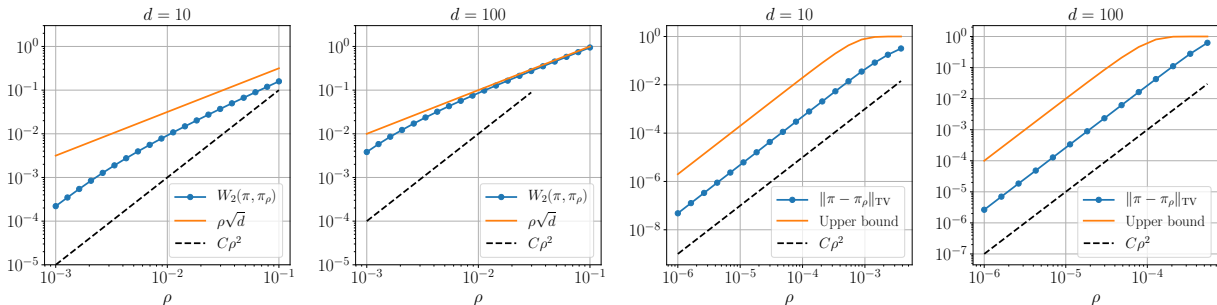


Figure 4: For  $d \in \{10, 100\}$ , illustration of the quantitative bounds (20) and (29) associated to 2-Wasserstein and total variation distances, respectively. The decay in  $\mathcal{O}(\rho^2)$  is shown via the dashed line  $C\rho^2$  where  $C$  is a constant.

## 5.2 Sparse linear regression

We study here a generalized version of the least absolute shrinkage and selection operator (lasso) regression problem analyzed by Park and Casella (2008). We assume a standard linear regression problem where centered observations  $\mathbf{y} \in \mathbb{R}^n$  are related to the unknown parameters  $\boldsymbol{\theta} \in \mathbb{R}^d$  via the model  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ , where  $\mathbf{X} \in \mathbb{R}^{n \times d}$  stands for a known standardized design matrix and  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$ . By considering a generalized Laplacian prior distribution for  $\boldsymbol{\theta}$ , the target posterior distribution has density for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ ,

$$\pi(\boldsymbol{\theta}) \triangleq \pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 - g(\mathbf{B}\boldsymbol{\theta})\right) \quad (44)$$

where  $g(\mathbf{B}\boldsymbol{\theta}) = \tau \|\mathbf{B}\boldsymbol{\theta}\|_1$  with  $\tau > 0$  and  $\mathbf{B} \in \mathbb{R}^{k \times d}$  an arbitrary matrix acting on  $\boldsymbol{\theta}$ . The choice of such a prior may promote a form of sparsity (lasso). For instance, this matrix  $\mathbf{B}$  might stand for a  $p$ -th order difference operator (Bredies et al. 2010) which is highly used



in signal and image processing problems. As an archetypal example, the case  $p = 1$  leads to the well-known total variation regularization function (Chambolle et al. 2010) used to recover piecewise constant signals.

Note that because of the presence of the matrix  $\mathbf{B}$ , finding an exact data augmentation leading to an efficient sampling scheme is not possible for the general case  $\mathbf{B} \neq \mathbf{I}_d$ . Instead, an AXDA model makes the posterior sampling task possible. Indeed, by regularizing the prior with a Gaussian term, the joint density  $\pi_\rho$  writes

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 - g(\mathbf{z}) - \frac{1}{2\rho^2} \|\mathbf{B}\mathbf{x} - \mathbf{z}\|_2^2\right). \quad (45)$$

By resorting to a Gibbs algorithm to sample from (45), one can now use a simple data augmentation scheme (Park and Casella 2008) to sample from the  $\mathbf{z}$ -conditional. On the other hand, sampling from the  $\boldsymbol{\theta}$ -conditional, which is a multivariate Gaussian distribution, can be undertaken efficiently with state-of-the-art approaches (Papandreou and Yuille 2010; Barbos et al. 2017; Marnissi et al. 2018).

In this specific case, the potential  $g_\rho$  associated to the smoothed prior distribution (see (35)) has a closed-form expression given for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ , by

$$\begin{aligned} g_\rho(\boldsymbol{\theta}) &= \frac{k}{2} \log(2\pi\rho^2) - \log \prod_{i=1}^k \int_{\mathbb{R}} \exp\left(-\tau|z_i| - \frac{1}{2\rho^2}(\mathbf{b}_i^T \boldsymbol{\theta} - z_i)^2\right) dz_i \\ &= \frac{k}{2} \log(2\pi\rho^2) \end{aligned} \quad (46)$$

$$- \log \prod_{i=1}^k \left( a(\boldsymbol{\theta}) \left[ \exp(b(\boldsymbol{\theta})^2) \{1 - \text{erf}(b(\boldsymbol{\theta}))\} + \exp(c(\boldsymbol{\theta})^2) \{1 - \text{erf}(c(\boldsymbol{\theta}))\} \right] \right) \quad (47)$$

with  $a(\boldsymbol{\theta}) = \sqrt{\pi\rho^2/2} \exp(-(\mathbf{b}_i^T \boldsymbol{\theta})^2/(2\rho^2))$ ,  $b(\boldsymbol{\theta}) = \sqrt{\rho^2/2}(\tau - \mathbf{b}_i^T \boldsymbol{\theta}/\rho^2)$ ,  $c(\boldsymbol{\theta}) = \sqrt{\rho^2/2}(\tau + \mathbf{b}_i^T \boldsymbol{\theta}/\rho^2)$  and  $\mathbf{b}_i \in \mathbb{R}^d$  standing for the  $i$ -th row of  $\mathbf{B}$ . Note that in more general cases where  $g_\rho$  has no closed form, one can estimate it by a Monte Carlo approximation.

Figure 5 shows the behavior of the regularized potential  $g_\rho$  defined in (47) for several values of the parameter  $\rho$  along with the associated smoothed prior and posterior distributions. For simplicity and pedagogical reasons, the univariate case corresponding to  $\boldsymbol{\theta} = \theta_1 \in \mathbb{R}$  and  $\mathbf{B} = 1$  has been considered. The regularization parameter  $\tau$  has been set to  $\tau = 1$ . The contours of the shaded area correspond to  $g + L_\rho$  and  $g + U_\rho$ . The potential  $g_\rho$  is a smooth approximation of the potential  $g$  associated to the initial prior as

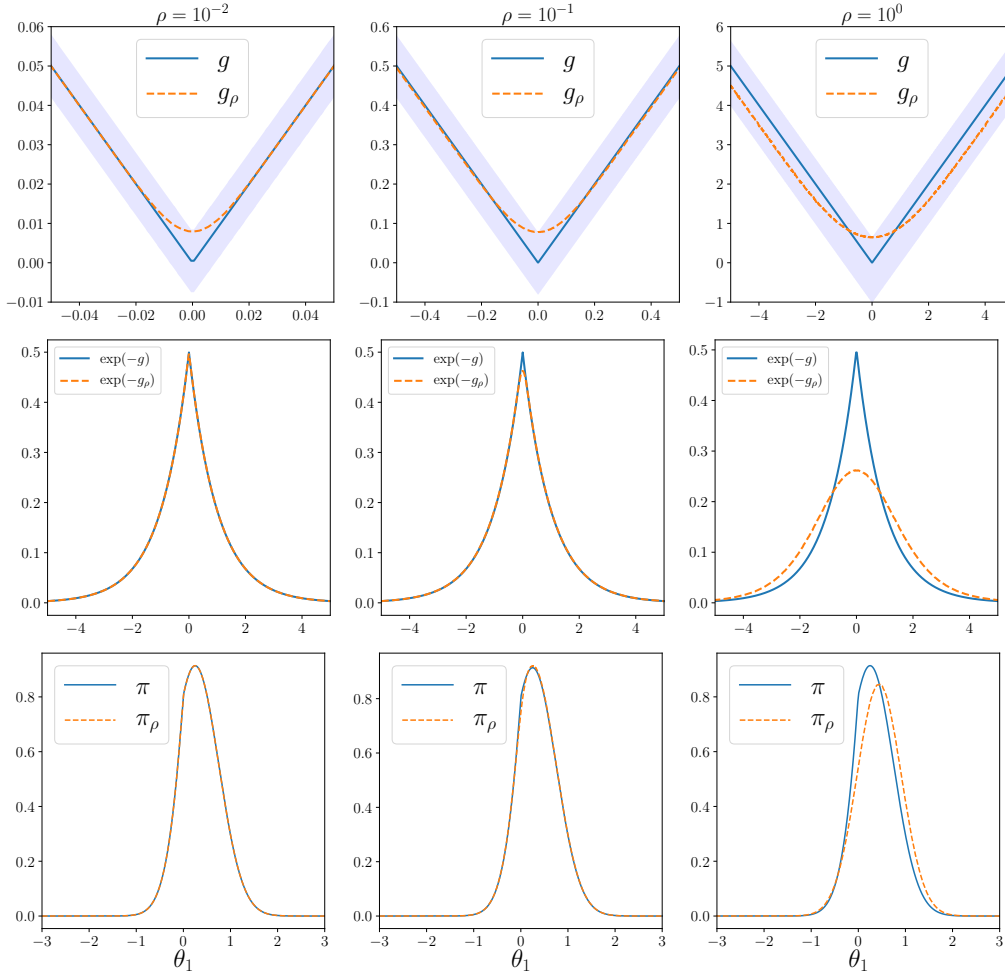


Figure 5: From left to right,  $\rho = 0.01$ ,  $\rho = 0.1$  and  $\rho = 1$ . (1st row) Behaviors of  $g$  (blue) and  $g_\rho$  (orange) where the contours of the shaded area correspond to  $g + L_\rho$  and  $g + U_\rho$ ; (2nd row) the corresponding normalized smoothed prior densities proportional to  $\exp(-g)$  and  $\exp(-g_\rho)$ ; (3rd row) posterior densities  $\pi_\rho$  w.r.t.  $\rho$ .

expected, see Property *iv*) in Proposition 1. Note that the inequalities derived in (36) are verified. Although this approximation seems similar to the Moreau-Yosida regularization of a non-smooth potential function (Combettes and Pesquet 2011), the rationale behind this approximation is different. Indeed, the Moreau-Yosida envelope stands for a particular instance of the infimal convolution between two convex functions (an initial potential and a Gaussian one). On the other hand,  $g_\rho$  is the potential associated to a smoothed density obtained by convolution with a Gaussian kernel. In addition, the third row of Figure 5

Table 2: Illustration of the bound derived in (40) for the marginal posterior  $\pi_\rho$  depicted in Section 5.2. The  $(1-\alpha)$ -credibility intervals  $\mathcal{C}_\alpha$  and  $\mathcal{C}_\alpha^\rho$  are the highest posterior density regions associated to each density with  $\alpha = 0.05$ .

$\rho$	$\mathcal{C}_\alpha$	$\mathcal{C}_\alpha^\rho$	$\int_{\mathcal{C}_\alpha^\rho} \pi(\theta_1) d\theta_1$	$\mathcal{I}_\alpha^\rho$
$10^{-3}$	[-0.47,1.24]	[-0.47,1.24]	0.95	[0.949,0.951]
$10^{-2}$	idem	[-0.47,1.24]	0.95	[0.948,0.952]
$10^{-1}$	idem	[-0.47,1.24]	0.95	[0.88,1]
$10^0$	idem	[-0.47,1.37]	0.96	[0.34,1]

shows the form of the posterior of  $\theta_1$  defined in (45) for  $y = 1$ ,  $x = 2$  and  $\sigma = 1$  and derived from the smoothed prior distributions shown in Figure 5. For sufficiently small values of  $\rho$ , the marginal  $\pi_\rho$  stands for a quite accurate approximation of the original target  $\pi$ .

Table 2 illustrates the bounds derived in (40) for  $\rho \in \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$ . For each case, the values of the bounds are summarized in the interval

$$\mathcal{I}_\alpha^\rho = [(1 - \alpha)N_\rho/D_{-d}(-L_f\rho), \min(1, (1 - \alpha)N_\rho/D_{-d}(L_f\rho))], \quad (48)$$

and the real coverage  $\int_{\mathcal{C}_\alpha^\rho} \pi(\theta_1) d\theta_1$  is also reported. The  $(1-\alpha)$ -credibility intervals  $\mathcal{C}_\alpha$  and  $\mathcal{C}_\alpha^\rho$  have been chosen to be the highest posterior density regions associated to each density with  $\alpha = 0.05$ . Note that the theoretical coverage interval  $\mathcal{I}_\alpha^\rho$  becomes informative only if  $\rho$  is sufficiently small which is not surprising since the assumptions on the potential of  $\pi_\rho$  are weak. Indeed, the form of the density (e.g. symmetry or unimodality) is not taken into account in the derived bounds. Regarding the empirical value of the coverage  $\int_{\mathcal{C}_\alpha^\rho} \pi(\theta_1) d\theta_1$ , we emphasize that the marginal  $\pi_\rho$  stands for a conservative approximation of  $\pi$  in this example. Indeed, in each case, the  $(1-\alpha)$ -credibility interval under  $\pi_\rho$  denoted  $\mathcal{C}_\alpha^\rho$  covers at least  $100(1 - \alpha)\%$  of the probability mass under  $\pi$ .

### 5.3 Illustration on an image inpainting problem

We illustrate here the correctness of the proposed approach on a multidimensional and non-Gaussian example which classically appears in image processing. To this purpose, we

consider the observation of a damaged and noisy image  $\mathbf{y} \in \mathbb{R}^n$  (represented as a vector by lexicographic ordering) related to the unknown original image  $\boldsymbol{\theta} \in \mathbb{R}^d$  by the linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \quad (49)$$

where  $n < d$ ,  $\mathbf{H} \in \mathbb{R}^{n \times d}$  stands for a decimation binary matrix. The dimension  $d$  being typically large (e.g.,  $10^3 \leq d \leq 10^9$ ), these problems require scalable inference algorithms. Since the matrix  $\mathbf{H}$  is not invertible, the linear inverse problem (49) is ill-posed. To cope with this issue, we assign the total variation prior distribution to the unknown parameter  $\boldsymbol{\theta}$ , leading to the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{H}\boldsymbol{\theta}\|_2^2 - \tau \sum_{1 \leq i \leq d} \|(\mathbf{D}\boldsymbol{\theta})_i\|_2\right), \quad (50)$$

where  $\tau > 0$  is a regularization parameter,  $\mathbf{D}\boldsymbol{\theta} = (\mathbf{D}_1\boldsymbol{\theta}, \mathbf{D}_2\boldsymbol{\theta}) \in \mathbb{R}^{2 \times d}$  is the two-dimensional discrete gradient associated to the image  $\boldsymbol{\theta}$  and the notation  $\mathbf{M}_i$  stands for the  $i$ -th column of the matrix  $\mathbf{M}$ , see [Chambolle et al. \(2010\)](#) for more details about the total variation regularization. The presence of the operator  $\mathbf{D}$  and the non-differentiability of the total variation norm rule out the use of common data augmentation schemes and simulation-based algorithms (e.g., Hamiltonian and Langevin Monte Carlo methods). Possible surrogates are proximal MCMC methods ([Pereyra 2016](#); [Durmus et al. 2018](#)) which replace the non-differentiable posterior distribution by a smooth approximation based on the proximity operator ([Combettes and Pesquet 2011](#)) of the total variation norm. However, the latter does not admit a closed-form expression and iterative routines are commonly used to approximate the latter ([Chambolle 2004](#)) leading to higher computational costs.

To mitigate these issues, we propose to rely on a particular instance of AXDA by smoothing the total variation prior with a Gaussian term, leading to the approximate joint posterior density

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{H}\boldsymbol{\theta}\|_2^2 - \tau \sum_{1 \leq i \leq d} \|\mathbf{z}_i\|_2 - \frac{1}{2\rho^2} \|\mathbf{z} - \mathbf{D}\boldsymbol{\theta}\|_2^2\right), \quad (51)$$

where  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^{2 \times d}$ . By relying on (51), the inference is now simplified and can be conducted with a Gibbs sampler, see Section 4 in the supplementary material. Since  $\ker(\mathbf{H}) \cap \ker(\mathbf{D}) = \{\mathbf{0}_d\}$ , the conditional posterior distribution of  $\boldsymbol{\theta}$  is a non-degenerate

multivariate Gaussian distribution. Samples from the latter can be obtained efficiently with the two-dimensional discrete Fourier transform by exploiting the periodic boundary conditions for  $\boldsymbol{\theta}$  (Wang et al. 2008; Marnissi et al. 2018). On the other hand, samples from  $\pi_\rho(\mathbf{Z}|\boldsymbol{\theta})$  can be drawn efficiently using exact data augmentation, see Kyung et al. (2010). All the inference details are given in Section 4 of the supplementary material.

We illustrate the proposed approximate model  $\pi_\rho$  by considering the Shepp-Logan phantom magnetic resonance image of size  $100 \times 100$  ( $d = 10^4$ ), see Figure 6. We artificially damaged and added noise to this image to build a noisy observation  $\mathbf{y}$  consisting of 90% randomly selected pixels of the initial image. The standard deviation of the Gaussian noise and the regularization parameter have been set to  $\sigma = 7 \times 10^{-2}$  (corresponding to a SNR of 58dB) and  $\tau = 5$ , respectively. The tolerance parameter has been set to  $\rho = 0.1$ .

In order to assess the bias of the proposed approach, we implemented the Moreau-Yosida Metropolis-adjusted Langevin algorithm (MYMALA) of Durmus et al. (2018), specifically designed to sample exactly from high-dimensional and non-smooth posterior distributions. For all the MCMC algorithms, the initialization has been set to  $\boldsymbol{\theta}^{[0]} = \mathbf{0}_d$ . We generated  $10^5$  samples and kept the last  $5 \times 10^4$  ones.

Figure 6 shows the minimum mean square estimate (MMSE) under  $\pi_\rho$  along with the original image. One can denote that the MMSE under  $\pi_\rho$  is visually similar to the original image and hence coherent with the reconstruction task. The relative residual error between the former and the MMSE under  $\pi$  is of order 2%. The main differences are located on the boundaries of the image, as depicted in the figure on the left which shows the absolute difference between the pixels of the two posterior means  $\mathbb{E}_\pi(\boldsymbol{\theta})$  and  $\mathbb{E}_{\pi_\rho}(\boldsymbol{\theta})$ .

To emphasize the correctness of the proposed approach beyond the comparison between pointwise estimates, we also paid attention to the comparison between posterior credibility sets induced by both  $\pi$  and  $\pi_\rho$ . To this purpose, we considered the highest posterior density region given by

$$\mathcal{C}_\alpha^* = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid f(\boldsymbol{\theta}) \leq \gamma_\alpha\}, \quad (52)$$

where  $\gamma_\alpha \in \mathbb{R}$  is such that  $\int_{\mathcal{C}_\alpha^*} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = 1 - \alpha$  and  $f$  is the potential function associated to  $\pi(\boldsymbol{\theta}|\mathbf{y})$ .

Figure 7 shows the different values of the scalar summary  $\gamma_\alpha$  estimated using  $\pi$  and

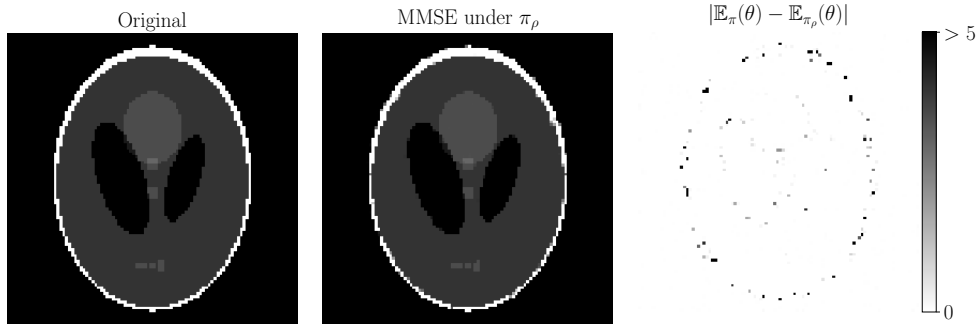


Figure 6: From left to right: original image, minimum mean square estimate (MMSE) under  $\pi_\rho$  and absolute bias between the posterior means under  $\pi_\rho$  and  $\pi$ .

the scalar  $\gamma_\alpha^\rho$  estimated using  $\pi_\rho$  for  $\alpha \in [0.01, 0.99]$ . Note that the approximation error associated to  $\gamma_\alpha$  is of order 2.6% whatever the value of  $\alpha$ , which supports the use of  $\pi_\rho$  to conduct Bayesian uncertainty analysis in this problem. After the burn-in period, the efficiency of the Gibbs algorithm used to sample from  $\pi_\rho$  has been measured by comparing the effective sample size (ESS) associated to the slowest component of  $\boldsymbol{\theta}$  to the one obtained with MYMALA. We found that the two ESS were roughly similar but the cost per iteration of the Gibbs sampler (0.079 sec/iteration) is almost two times lower than that of MYMALA (0.144 sec/iteration)<sup>1</sup>. In addition, the number of iterations required to reach high-probability regions is much less important for the Gibbs sampler than for MYMALA, showing the interest of AXDA, see Figure 7.

## 6 Conclusion

This paper presented a unifying framework for asymptotically exact data augmentation (AXDA) schemes. AXDA introduces approximate densities in order to simplify the inference. By building on existing works which considered special instances of AXDA, we illustrate potential benefits that can be inherited by the proposed framework such as distributed computations, robustness or sophisticated inference schemes from the ABC literature. On top of these qualitative properties, we derived a set of theoretical guarantees on the bias involved in the proposed methodology. The latter encompass a large class of AXDA mod-

<sup>1</sup>Both algorithms have been implemented in MATLAB with the same level of efficiency.

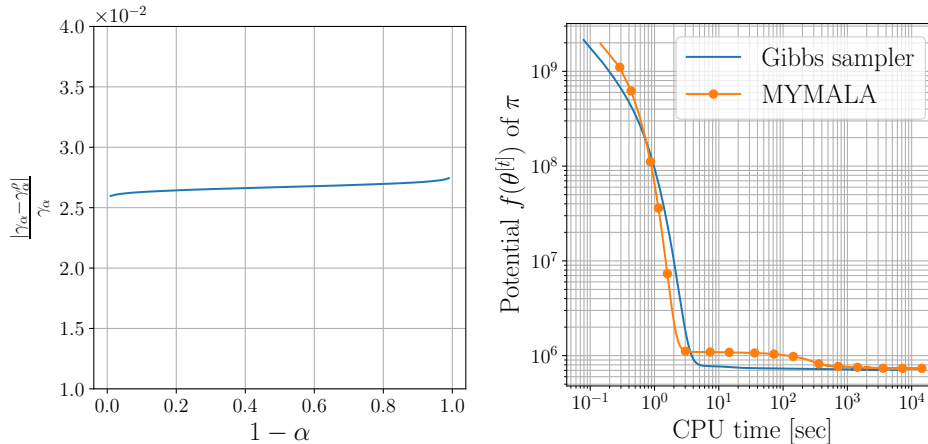


Figure 7: (left) relative error between the threshold value estimated with  $\pi$  denoted  $\gamma_\alpha$  and the one estimated with  $\pi_\rho$  denoted  $\gamma_\alpha^\rho$  and (right) Potential  $f = -\log \pi$  w.r.t. the number of iterations  $t$  for both MYMALA and the Gibbs sampler targeting  $\pi_\rho$ .

els and a detailed non-asymptotic analysis has been done for Gaussian smoothing. These results have been illustrated on several cases that can arise in statistical learning or signal processing showing the broad scope of application of the proposed approach. In practice, we emphasize that AXDA models can remarkably improve the inference task in big data and high-dimensional settings. In summary, at the price of an approximation which comes with theoretical guarantees, AXDA approaches appear to be a general, systematic and efficient way to conduct simple inference in a wide variety of large-scale problems. They provide accurate estimates with relevant confidence intervals that are crucial in many applications, in particular when no ground truth is available.

## Acknowledgements

Part of this work has been supported by the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) under grant agreement ANITI ANR-19-PI3A-0004.

## SUPPLEMENTARY MATERIALS

**Appendices:** The supplementary material includes the proofs of Theorems 1 and 2; Propositions 1, 2, 3, 4 and 5; and Corollary 1 and 3. It also includes additional details about

standard kernels and Bregman divergences, inference details associated to the image inpainting example in Section 5.3 and derivations of classical inference algorithms to target AXDA models. (supplementary\_material.pdf, pdf file)

**Package for AXDA:** The computer code associated to the illustrations and experiments described in this paper is also available online. More precisely, the package “AXDA” contains a PYTHON jupyter notebook to reproduce all the tables and figures of the paper and a MATLAB-code associated to the image inpainting example along with a README file. (axda.zip, zipped file)



# SUPPLEMENTARY MATERIALS

## 1 Proofs

### 1.1 Proof of Proposition 1

Property *i*) follows from the fact that  $\pi_\rho$  stands for a convolution integral between  $\pi$  and  $\kappa_\rho$ , i.e.  $\pi_\rho = \pi * \kappa_\rho$ . Therefore, the expectation and variance under  $\pi_\rho$  are the sum of the expectations and variances of two independent random variables under  $\pi$  and  $\kappa_\rho$  respectively. Property *ii*) follows directly from (Folland 1999, Proposition 8.6). Property *iii*) follows from the fact that log-concavity is preserved by convolution of distributions (Dharmadhikari and Joag-Dev 1988, Theorem 2.18). Finally, Property *iv*) follows from the dominated convergence theorem since  $\pi \in L^1$ ,  $\kappa_\rho \in \mathcal{C}^\infty(\mathbb{R}^d)$  and for all  $k \geq 0$ ,  $|\partial^k \kappa_\rho| \leq C_k$  (Folland 1999, Proposition 8.10).

### 1.2 Proof of Proposition 2

The proof can be found in Ambrosio et al. (2008, Lemma 7.1.10). Since it is quite short, we recall it hereafter for completeness. We have

$$W_p^p(\pi, \pi_\rho) = \min_{\mu} \left\{ \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|_2^p d\mu(\boldsymbol{\theta}, \mathbf{z}); \mu \in \Gamma(\pi_\rho, \pi) \right\} \quad (53)$$

$$\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|_2^p \pi_\rho(\boldsymbol{\theta}, \mathbf{z}) d\boldsymbol{\theta} d\mathbf{z} \quad (54)$$

$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|_2^p \kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \pi(\mathbf{z}) d\boldsymbol{\theta} d\mathbf{z} \quad (55)$$

$$= \rho^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|_2^p K(\rho^{-1}(\boldsymbol{\theta} - \mathbf{z})) \pi(\mathbf{z}) d\boldsymbol{\theta} d\mathbf{z} \quad (56)$$

$$= \rho^p \int_{\mathbb{R}^d} \|\mathbf{u}\|_2^p K(\mathbf{u}) d\mathbf{u} \int_{\mathbb{R}^d} \pi(\mathbf{z}) d\mathbf{z} \quad (57)$$

$$= \rho^p \int_{\mathbb{R}^d} \|\mathbf{u}\|_2^p K(\mathbf{u}) d\mathbf{u}. \quad (58)$$

### 1.3 Proof of Proposition 3

Let  $\boldsymbol{\theta} \in \mathbb{R}^d$ . Since  $\pi$  has been assumed to be analytic and twice differentiable with  $\mathbf{H}_\pi$  being continuous, there exists  $\tilde{\boldsymbol{\theta}}$  lying between  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}$  such that

$$\pi_\rho(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} \pi(\mathbf{z}) \kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} \quad (59)$$

$$= \frac{\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}) \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}} \quad (60)$$

$$= \frac{\int_{\mathbb{R}^d} \left[ \pi(\boldsymbol{\theta}) - \sqrt{\rho} \nabla \pi(\boldsymbol{\theta})^T \mathbf{u} + \frac{\rho}{2} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}}) \mathbf{u} \right] \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}}, \quad (61)$$

where  $\mathbf{H}_\pi$  stands for the Hessian matrix of  $\pi$ .

It follows that

$$\pi_\rho(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \quad (62)$$

$$- \sqrt{\rho} \nabla \pi(\boldsymbol{\theta})^T \int_{\mathbb{R}^d} \mathbf{u} \frac{\exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}} \quad (63)$$

$$+ \frac{\rho}{2} \frac{\int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}}) \mathbf{u} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) d\mathbf{u}} \quad (64)$$

We now show that (63) =  $O(\sqrt{\rho})$  and (64) =  $\mathcal{O}(\rho)$ . To this purpose, we use the analyticity and two times differentiability of  $d_\psi$  w.r.t. to its first argument and the continuity of  $\mathbf{H}_{d_\psi}$ . By definition of the Bregman divergence (see Definition 1 in the main paper),  $d_\psi(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$  and  $\nabla_{\mathbf{z}} d_\psi(\mathbf{z}, \boldsymbol{\theta}) \Big|_{\mathbf{z}=\boldsymbol{\theta}} = \mathbf{0}_d$  so that, for all  $\mathbf{u} \in \mathbb{R}^d$ ,

$$d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta}) = \frac{\rho}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}, \quad (65)$$

where  $\boldsymbol{\theta}'$  lies between  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}$ .

We first prove (64) =  $\mathcal{O}(\rho)$ . Using (65), we can re-write (64) as

$$(12) = \frac{\rho \int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \mathrm{d}\mathbf{u}}{2 \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \mathrm{d}\mathbf{u}}. \quad (66)$$

Since  $\lim_{\rho \rightarrow 0} \boldsymbol{\theta}' = \boldsymbol{\theta}$  and  $\lim_{\rho \rightarrow 0} \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$ , we will use the dominated convergence theorem using that

$$\lim_{\rho \rightarrow 0} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) = \mathbf{u}^T \mathbf{H}_\pi(\boldsymbol{\theta}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}) \mathbf{u}\right). \quad (67)$$

In addition, since  $d_\psi$  is strictly convex w.r.t. its first argument,  $\mathbf{H}_{d_\psi}$  is a symmetric and positive-definite matrix. By using that  $\|\mathbf{H}_\pi\| \leq C < \infty$  and  $\|\mathbf{H}_{d_\psi}\| \geq c > 0$ , we have:

$$\left| \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \right| \leq C \|\mathbf{u}\|_2^2 \exp\left(-\frac{c}{2} \mathbf{u}^T \mathbf{u}\right), \quad (68)$$

which is integrable on  $\mathbb{R}^d$ . From the dominated convergence theorem, it follows that

$$\int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \mathrm{d}\mathbf{u} = \int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\boldsymbol{\theta}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}) \mathbf{u}\right) \mathrm{d}\mathbf{u} + o(1). \quad (69)$$

Similarly,

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \mathrm{d}\mathbf{u} = \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}) \mathbf{u}\right) \mathrm{d}\mathbf{u} + o(1). \quad (70)$$

Hence,

$$(12) = \frac{\rho \int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\boldsymbol{\theta}) \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}) \mathbf{u}\right) \mathrm{d}\mathbf{u}}{2 \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}) \mathbf{u}\right) \mathrm{d}\mathbf{u}} + o(\rho) \quad (71)$$

$$= \frac{\rho}{2} \text{Trace}\left(\mathbf{H}_\pi(\boldsymbol{\theta}) \mathbf{H}_{d_\psi}(\boldsymbol{\theta})^{-1}\right) + o(\rho). \quad (72)$$

We now prove (63) =  $\mathcal{O}(\sqrt{\rho})$ . Using (65), it follows that

$$(11) = -\sqrt{\rho} \nabla \pi(\boldsymbol{\theta})^T \frac{\int_{\mathbb{R}^d} \mathbf{u} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \mathrm{d}\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}') \mathbf{u}\right) \mathrm{d}\mathbf{u}}. \quad (73)$$

Again, since  $\|H_{d_\psi}\|$  has been assumed to be lower bounded, it follows from the dominated convergence theorem that

$$(11) = -\sqrt{\rho}\nabla\pi(\boldsymbol{\theta})^T \frac{\int_{\mathbb{R}^d} \mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) d\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) d\mathbf{u}} + o(\sqrt{\rho}) \quad (74)$$

$$= -\sqrt{\rho}\nabla\pi(\boldsymbol{\theta})^T \frac{\int_{\mathbb{R}^d} \mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) d\mathbf{u}}{\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) d\mathbf{u}} + o(\sqrt{\rho}) \quad (75)$$

$$= o(\sqrt{\rho}) = \mathcal{O}(\sqrt{\rho}). \quad (76)$$

## 1.4 Proof of Theorem 1

We are interested in controlling w.r.t.  $\rho$  the quantity  $\|\pi_\rho - \pi\|_{\text{TV}}$ . To this purpose, let assume that  $f$  satisfies  $(A_1)$  in the main paper. In the following, we will assume for simplicity reasons that  $\pi$  stands for a pdf associated to the random variable  $\boldsymbol{\theta}$ . The case when  $\pi$  is a likelihood is treated right after. Under this convention, it follows

$$\begin{aligned} \|\pi_\rho - \pi\|_{\text{TV}} &= \frac{1}{2} \int_{\mathbb{R}^d} |\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta})| d\boldsymbol{\theta} \\ &= \frac{1}{2} \int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left| \frac{C_\pi}{C_{\pi_\rho}} \mathcal{K}(\boldsymbol{\theta}) - 1 \right| d\boldsymbol{\theta}, \end{aligned} \quad (77)$$

where  $C_\pi$  and  $C_{\pi_\rho}$  are the normalizing constants associated to  $\pi$  and  $\pi_\rho$ , respectively, and

$$\mathcal{K}(\boldsymbol{\theta}) = \frac{\pi_\rho(\boldsymbol{\theta})C_{\pi_\rho}}{\pi(\boldsymbol{\theta})C_\pi} \quad (78)$$

$$= \int_{\mathbb{R}^d} \exp\left(f(\boldsymbol{\theta}) - f(\mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z}. \quad (79)$$

Note that

$$\int_{\mathbb{R}^d} \mathcal{K}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{C_{\pi_\rho}}{C_\pi}. \quad (80)$$

Since  $f$  is assumed to be  $L_f$ -Lipschitz, we have

$$\mathcal{K}(\boldsymbol{\theta}) \leq \int_{\mathbb{R}^d} \exp\left(L_f \|\boldsymbol{\theta} - \mathbf{z}\|_2 - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z}. \quad (81)$$

We make the change of variables  $\mathbf{u} = \mathbf{z} - \boldsymbol{\theta}$ , which leads to

$$\mathcal{K}(\boldsymbol{\theta}) \leq \int_{\mathbb{R}^d} \exp\left(L_f \|\mathbf{u}\|_2 - \frac{1}{2\rho^2} \|\mathbf{u}\|_2^2\right) d\mathbf{u}. \quad (82)$$

Then, with another change of variables  $t = \|\mathbf{u}\|_2$ , it follows

$$\mathcal{K}(\boldsymbol{\theta}) \leq \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty t^{d-1} \exp\left(L_f t - \frac{1}{2\rho^2} t^2\right) dt. \quad (83)$$

This integral admits a closed-form expression ([Gradshteyn and Ryzhik 2015](#), Formula 3.462 1.) by introducing the special parabolic cylinder function  $D_{-d}$  defined for all  $d > 0$  and  $z \in \mathbb{R}$  by

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz-x^2/2} x^{d-1} dx. \quad (84)$$

Then,

$$\mathcal{K}(\boldsymbol{\theta}) \leq A(\rho), \quad (85)$$

where

$$A(\rho) = \frac{2\pi^{d/2} \rho^d \Gamma(d) \exp\left(\frac{L_f^2 \rho^2}{4}\right)}{\Gamma\left(\frac{d}{2}\right)} D_{-d}(-L_f \rho). \quad (86)$$

Then, with (80) and (85), we also have

$$\frac{C_\pi}{C_{\pi_\rho}} \geq \frac{1}{A(\rho)}. \quad (87)$$

We now use the triangle inequality in (77) which leads to

$$\begin{aligned} \|\pi_\rho - \pi\|_{\text{TV}} &\leq \frac{1}{2} \left( \int_{\mathbb{R}^d} \left| \frac{C_\pi}{C_{\pi_\rho}} \mathcal{K}(\boldsymbol{\theta}) - \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) \right| \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} + \int_{\mathbb{R}^d} \left| \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) - 1 \right| \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \\ &= \frac{1}{2} \left( \int_{\mathbb{R}^d} \left( \frac{C_\pi}{C_{\pi_\rho}} \mathcal{K}(\boldsymbol{\theta}) - \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) \right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} + \int_{\mathbb{R}^d} \left( 1 - \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) \right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right). \end{aligned} \quad (88)$$

The first term in this upper bound writes

$$\begin{aligned} \int_{\mathbb{R}^d} \left( \frac{C_\pi}{C_{\pi_\rho}} - \frac{1}{A(\rho)} \right) \mathcal{K}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} &= 1 - \frac{1}{A(\rho)} \int_{\mathbb{R}^d} \mathcal{K}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\mathbb{R}^d} \left( 1 - \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) \right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \quad (89)$$

This allows us to bound (88), that is

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq \int_{\mathbb{R}^d} \left( 1 - \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) \right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (90)$$

Using one more time the  $L_f$ -Lipschitz assumption on  $f$ , we have for all  $\boldsymbol{\theta}, \mathbf{z}$ ,

$$-(f(\mathbf{z}) - f(\boldsymbol{\theta})) \geq -|f(\mathbf{z}) - f(\boldsymbol{\theta})| \geq -L_f \|\boldsymbol{\theta} - \mathbf{z}\|_2, \quad (91)$$

$$\text{so that } \mathcal{K}(\boldsymbol{\theta}) \geq \int_{\mathbb{R}^d} \exp\left(-L_f \|\boldsymbol{\theta} - \mathbf{z}\|_2 - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z}. \quad (92)$$

With the same changes of variables as above, it follows

$$\mathcal{K}(\boldsymbol{\theta}) \geq B(\rho), \quad (93)$$

where

$$B(\rho) = \frac{2\alpha_{\text{Vol}} \pi^{d/2} \rho^d \Gamma(d) \exp\left(\frac{L_f^2 \rho^2}{4}\right)}{\Gamma\left(\frac{d}{2}\right)} D_{-d}(L_f \rho). \quad (94)$$

Then we have  $1 - \frac{1}{A(\rho)} \mathcal{K}(\boldsymbol{\theta}) \leq 1 - \frac{B(\rho)}{A(\rho)}$  which combined with (90) yields

$$\|\pi_\rho - \pi\|_{\text{TV}} \leq 1 - \frac{D_{-d}(L_f \rho)}{D_{-d}(-L_f \rho)}. \quad (95)$$

**Note:** When  $\pi = \pi(\mathbf{y}|\boldsymbol{\theta})$  is a likelihood, (77) becomes

$$\|\pi_\rho - \pi\|_{\text{TV}} = \frac{1}{2} \int_{\mathbb{R}^n} \pi(\mathbf{y}|\boldsymbol{\theta}) \left| \frac{C_\pi}{C_{\pi_\rho}} \mathcal{K}(\mathbf{y}; \boldsymbol{\theta}) - 1 \right| d\mathbf{y}. \quad (96)$$

Since  $L_f$  is assumed to be independent of  $\mathbf{y}$ , the same type of proof can be followed in this case and yields the same quantitative bound.

## 1.5 Proof of Corollary 1

The parabolic cylinder function when  $d > 0$  has the following expression ([Gradshteyn and Ryzhik 2015](#), Formula 9.241 2.)

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz-x^2/2} x^{d-1} dx. \quad (97)$$

In the limiting case when  $z \rightarrow 0$ , a first order Taylor expansion of  $e^{-xz}$  gives

$$\begin{aligned} D_{-d}(z) &= \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-x^2/2} x^{d-1} (1 - xz + o(z)) dx \\ &= \frac{\exp(-z^2/4)}{\Gamma(d)} \left( \int_0^{+\infty} e^{-x^2/2} x^{d-1} dx - z \int_0^{+\infty} e^{-x^2/2} x^d dx + o(z) \right) \\ &= \frac{\exp(-z^2/4)}{\Gamma(d)} \left( \Gamma\left(\frac{d}{2}\right) 2^{d/2-1} - z \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(z) \right), \end{aligned} \quad (98)$$

recording that  $\int_0^{+\infty} e^{-x^2/2} x^d dx = \Gamma((d+1)/2) 2^{d/2-1/2}$  ([Gradshteyn and Ryzhik 2015](#), Formula 3.383 11.). Using (98) for  $z = \pm \rho L_f$  yields

$$\begin{aligned} 1 - \frac{D_{-d}(L_f \rho)}{D_{-d}(-L_f \rho)} &= 1 - \frac{\frac{\exp(-(\rho L_f)^2/4)}{\Gamma(d)} \left( \Gamma\left(\frac{d}{2}\right) 2^{d/2-1} - \rho L_f \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(\rho) \right)}{\frac{\exp(-(\rho L_f)^2/4)}{\Gamma(d)} \left( \Gamma\left(\frac{d}{2}\right) 2^{d/2-1} + \rho L_f \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(\rho) \right)} \\ &= 1 - \frac{\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} - \rho L_f \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(\rho)}{\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} \left( 1 + \rho \frac{L_f \Gamma\left(\frac{d+1}{2}\right) \sqrt{2}}{\Gamma\left(\frac{d}{2}\right)} + o(\rho) \right)} \\ &= 1 - \left( 1 - \rho \frac{L_f \Gamma\left(\frac{d+1}{2}\right) \sqrt{2}}{\Gamma\left(\frac{d}{2}\right)} + o(\rho) \right) \left( 1 - \rho \frac{L_f \Gamma\left(\frac{d+1}{2}\right) \sqrt{2}}{\Gamma\left(\frac{d}{2}\right)} + o(\rho) \right) \\ &= \frac{2\sqrt{2} \Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L_f \rho + o(\rho). \end{aligned} \quad (99)$$

## 1.6 Dependence of (99) with respect to the dimension

The gamma function  $\Gamma$  can be expressed for all  $z > 0$  as  $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx$ . When  $z$  is large, Stirling-like approximations give the following equivalent for  $\Gamma(z + 1/2)$  and  $\Gamma(z)$ :

$$\Gamma(z + 1/2) \underset{z \rightarrow +\infty}{\sim} \sqrt{2\pi} z^z e^{-z} \quad (100)$$

$$\Gamma(z) \underset{z \rightarrow +\infty}{\sim} \sqrt{2\pi} z^{z-1/2} e^{-z}. \quad (101)$$

So that when  $d$  is large

$$\begin{aligned} \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L_f \rho &\underset{d \rightarrow +\infty}{\sim} \frac{2\sqrt{2}\sqrt{2\pi}(d/2)^{d/2} e^{-d/2}}{\sqrt{2\pi}(d/2)^{d/2-1/2} e^{-d/2}} L_f \rho \\ &\underset{d \rightarrow +\infty}{\sim} 2\sqrt{2}(d/2)^{1/2} L_f \rho \\ &\underset{d \rightarrow +\infty}{\sim} 2L_f \rho d^{1/2}. \end{aligned} \quad (102)$$

## 1.7 Proof of Theorem 2

We now prove another bound on the TV distance when  $f$  satisfies  $(A_2)$ ,  $(A_3)$  and  $(A_4)$  in the main paper. The beginning of the proof follows the same lines as in the proof of Theorem 1 above when  $\pi$  stands for a pdf associated to  $\boldsymbol{\theta}$ . Hence, we have from (77) that

$$\|\pi_\rho - \pi\|_{\text{TV}} = \frac{1}{2} \int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left| 1 - \mathcal{K}(\boldsymbol{\theta}) \frac{C_\pi}{C_{\pi_\rho}} \right| d\boldsymbol{\theta}. \quad (103)$$

We now use the convexity of  $f$  to write for all  $\boldsymbol{\theta} \in \mathbb{R}^d, \mathbf{z} \in \mathbb{R}^d$ ,

$$f(\boldsymbol{\theta}) - f(\mathbf{z}) \leq \nabla f(\boldsymbol{\theta})^T (\boldsymbol{\theta} - \mathbf{z}). \quad (104)$$

By using (104) and (79), it follows that

$$\begin{aligned} \mathcal{K}(\boldsymbol{\theta}) &\leq \int_{\mathbb{R}^d} \exp\left(\nabla f(\boldsymbol{\theta})^T (\boldsymbol{\theta} - \mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z} \\ &= \exp\left(\frac{\rho^2}{2} \|\nabla f(\boldsymbol{\theta})\|_2^2\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\rho^2} \|\mathbf{z} - \boldsymbol{\theta} - \rho^2 \nabla f(\boldsymbol{\theta})\|_2^2\right) d\mathbf{z} \\ &= \exp\left(\frac{\rho^2}{2} \|\nabla f(\boldsymbol{\theta})\|_2^2\right) (2\pi\rho^2)^{d/2} = B_1(\boldsymbol{\theta}). \end{aligned} \quad (105)$$



By using again the convexity of  $f$ , we also have for all  $\boldsymbol{\theta} \in \mathbb{R}^d, \mathbf{z} \in \mathbb{R}^d$ ,

$$f(\boldsymbol{\theta}) - f(\mathbf{z}) \geq \nabla f(\mathbf{z})^T(\boldsymbol{\theta} - \mathbf{z}). \quad (106)$$

Then, (106) leads to

$$\begin{aligned} \mathcal{K}(\boldsymbol{\theta}) &\geq \int_{\mathbb{R}^d} \exp\left(\nabla f(\mathbf{z})^T(\boldsymbol{\theta} - \mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z} \\ &= \int_{\mathbb{R}^d} \exp\left(\nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta} - \mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) \\ &\quad \times \exp\left(-(\nabla f(\boldsymbol{\theta}) - \nabla f(\mathbf{z}))^T(\boldsymbol{\theta} - \mathbf{z})\right) d\mathbf{z}. \end{aligned} \quad (107)$$

We now use  $(A_2)$  in the main paper which leads to

$$\begin{aligned} \mathcal{K}(\boldsymbol{\theta}) &\geq \int_{\mathbb{R}^d} \exp\left(\nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta} - \mathbf{z}) - \left(\frac{1 + 2\rho^2 M_f}{2\rho^2}\right) \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z} \\ &= \exp\left(\frac{\rho^2}{2(1 + 2\rho^2 M_f)} \|\nabla f(\boldsymbol{\theta})\|_2^2\right) \left(\frac{2\pi\rho^2}{1 + 2\rho^2 M_f}\right)^{d/2} = B_2(\boldsymbol{\theta}). \end{aligned} \quad (108)$$

We now apply the triangle inequality in (103) which yields

$$\begin{aligned} \|\pi_\rho - \pi\|_{\text{TV}} &\leq \frac{1}{2} \int_{\mathbb{R}^d} \left| \frac{C_\pi}{C_{\pi_\rho}} - \frac{1}{B_1(\boldsymbol{\theta})} \right| \mathcal{K}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} + \frac{1}{2} \int_{\mathbb{R}^d} \left| \frac{\mathcal{K}(\boldsymbol{\theta})}{B_1(\boldsymbol{\theta})} - 1 \right| \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \frac{1}{2} \int_{\mathbb{R}^d} \left| \frac{C_\pi}{C_{\pi_\rho}} - \frac{1}{B_1(\boldsymbol{\theta})} \right| \mathcal{K}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} + \frac{1}{2} \int_{\mathbb{R}^d} \left(1 - \frac{\mathcal{K}(\boldsymbol{\theta})}{B_1(\boldsymbol{\theta})}\right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \quad (109)$$

The absolute value in the first term of (109) can be removed by noting that

$$\begin{aligned} \frac{C_\pi}{C_{\pi_\rho}} &= \frac{\int_{\mathbb{R}^d} \exp(-f(\boldsymbol{\theta})) d\boldsymbol{\theta}}{\int_{\mathbb{R}^d} \exp(-f(\mathbf{z})) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\rho^2} \|\mathbf{z} - \boldsymbol{\theta}\|_2^2\right) d\boldsymbol{\theta} d\mathbf{z}} \\ &\geq \frac{\int_{\mathbb{R}^d} \exp(-f(\boldsymbol{\theta})) d\boldsymbol{\theta}}{\int_{\mathbb{R}^d} \exp(-f(\mathbf{z})) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\rho^2} \|\mathbf{z} - \boldsymbol{\theta}\|_2^2\right) d\boldsymbol{\theta} d\mathbf{z}} \\ &= (2\pi\rho^2)^{-d/2} \\ &= \frac{\exp\left(\frac{\rho^2}{2} \|\nabla f(\boldsymbol{\theta})\|_2^2\right)}{B_1(\boldsymbol{\theta})} \\ &\geq \frac{1}{B_1(\boldsymbol{\theta})}. \end{aligned} \quad (110)$$

Then (109) becomes

$$\begin{aligned} \|\pi_\rho - \pi\|_{\text{TV}} &\leq \int_{\mathbb{R}^d} \left(1 - \frac{\mathcal{K}(\boldsymbol{\theta})}{B_1(\boldsymbol{\theta})}\right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\leq \int_{\mathbb{R}^d} \left(1 - \frac{B_2(\boldsymbol{\theta})}{B_1(\boldsymbol{\theta})}\right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \quad (111)$$

We now use the fact that  $-\exp(-u) \leq u - 1$  for all  $u \geq 0$  which yields

$$\begin{aligned} \|\pi_\rho - \pi\|_{\text{TV}} &\leq 1 + (1 + 2\rho^2 M_f)^{-d/2} \int_{\mathbb{R}^d} \left( \frac{\rho^4 M_f \|\nabla f(\boldsymbol{\theta})\|^2}{1 + 2\rho^2 M_f} - 1 \right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ (\text{with } (A_3)) &= 1 - (1 + 2\rho^2 M_f)^{-d/2} \left( 1 - \frac{\rho^4 M_f M_f}{1 + 2\rho^2 M_f} \right). \end{aligned} \quad (112)$$

The result in Corollary 3 in the main paper comes from a straightforward Taylor expansion of (112).

## 1.8 Proof of Corollary 3

Equation (79) becomes

$$\mathcal{K}(\boldsymbol{\theta}) = \prod_{j=1}^J \int_{\mathbb{R}^d} \exp\left(f_j(\boldsymbol{\theta}) - f_j(\mathbf{z}_j) - \frac{1}{2\rho_j^2} \|\boldsymbol{\theta} - \mathbf{z}_j\|_2^2\right) d\mathbf{z}_j = \prod_{j=1}^J \mathcal{K}_j(\boldsymbol{\theta}). \quad (113)$$

Bounding each term in (113) and following the proof of Theorem 1 detailed above completes the proof.

## 1.9 Proof of Proposition 4

By using (85) and (93) we have for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ ,

$$\begin{aligned} B(\rho) &\leq \int_{\mathbb{R}^d} \exp\left(f(\boldsymbol{\theta}) - f(\mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z} \leq A(\rho) \\ B(\rho) \exp(-f(\boldsymbol{\theta})) &\leq \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z} \leq A(\rho) \exp(-f(\boldsymbol{\theta})) \\ -\log A(\rho) + f(\boldsymbol{\theta}) &\leq -\log \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right) d\mathbf{z} \leq -\log B(\rho) + f(\boldsymbol{\theta}) \end{aligned}$$

So that

$$-\log A(\rho) + \frac{d}{2} \log(2\pi\rho^2) \leq f_\rho(\boldsymbol{\theta}) - f(\boldsymbol{\theta}) \leq -\log B(\rho) + \frac{d}{2} \log(2\pi\rho^2). \quad (114)$$

The result of Proposition 4 follows from the definition of  $A(\rho)$  and  $B(\rho)$ .

## 1.10 Proof of Proposition 5

By using (85) and (93) it follows, for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ ,

$$B(\rho) \leq \mathcal{K}(\boldsymbol{\theta}) \leq A(\rho) \tag{115}$$

$$B(\rho)C_\pi\pi(\boldsymbol{\theta}) \leq \mathcal{K}(\boldsymbol{\theta})C_\pi\pi(\boldsymbol{\theta}) \leq A(\rho)C_\pi\pi(\boldsymbol{\theta}). \tag{116}$$

Using (79) yields

$$B(\rho)\pi(\boldsymbol{\theta}) \leq \pi_\rho(\boldsymbol{\theta})\frac{C_{\pi_\rho}}{C_\pi} \leq A(\rho)\pi(\boldsymbol{\theta}) \tag{117}$$

$$B(\rho)\pi(\boldsymbol{\theta}) \leq \pi_\rho(\boldsymbol{\theta})(2\pi\rho^2)^{d/2} \leq A(\rho)\pi(\boldsymbol{\theta}). \tag{118}$$

Using (86) and (94) gives

$$\frac{N_\rho}{D_{-d}(-L_f\rho)}\pi_\rho(\boldsymbol{\theta}) \leq \pi(\boldsymbol{\theta}) \leq \frac{N_\rho}{D_{-d}(L_f\rho)}\pi_\rho(\boldsymbol{\theta}), \tag{119}$$

where the constant  $N_\rho$  has been defined in (39) in the main paper.

Let  $\mathcal{C}_\alpha^\rho$  an arbitrary  $(1 - \alpha)$ -credibility region under  $\pi_\rho$ . By integrating (119) on  $\mathcal{C}_\alpha^\rho$ ,

$$\frac{N_\rho}{D_{-d}(-L_f\rho)}(1 - \alpha) \leq \int_{\mathcal{C}_\alpha^\rho} \pi(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \frac{N_\rho}{D_{-d}(L_f\rho)}(1 - \alpha). \tag{120}$$

Since  $\mathcal{C}_\alpha^\rho \subseteq \mathbb{R}^d$  and  $\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$ , the upper bound in (120) can be replaced by

$$\min \left\{ 1, \frac{N_\rho}{D_{-d}(L_f\rho)}(1 - \alpha) \right\}.$$

## 1.11 Lipschitz loss functions - Dependence w.r.t. the number of observations

Combining Corollary 4 in the main paper with (99), we have:

$$\|\pi - \pi_\rho\|_{\text{TV}} \leq 1 - \prod_{i=1}^n \left( 1 - \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L_{f_i} \rho + o(\rho) \right) \quad (121)$$

$$= \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \rho \sum_{i=1}^n L_{f_i} + o(\rho) \quad (122)$$

$$\leq \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right) \max_{i \in [n]} L_{f_i}}{\Gamma\left(\frac{d}{2}\right)} n\rho + o(\rho). \quad (123)$$

## 2 On the choice of $\kappa_\rho$

Section 2.2 of the main document provides two distinct ways to choose  $\kappa_\rho$ . We can first consider a kernel  $K$  that is a positive function such that  $\int_{\mathbb{R}^d} K(\mathbf{u})d\mathbf{u} = 1$  and  $K(-\mathbf{u}) = K(\mathbf{u})$ , for all  $\mathbf{u} \in \mathbb{R}^d$ . Based on the latter, we define for all  $\mathbf{z}, \boldsymbol{\theta} \in \Theta$ ,  $\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto_{\mathbf{z}} \rho^{-d} K(\rho^{-1}(\boldsymbol{\theta} - \mathbf{z}))$  (Dang and Ehrhardt 2012). Table 3 lists some classical examples of symmetric kernels  $K(\cdot)$  which are not necessarily compactly supported. For sake of simplicity, we only define univariate versions of them but they can obviously be generalized in higher dimension. Figure 8 illustrates these kernels. Some of them have for instance been used in ABC approaches (Sisson et al. 2018b).

Table 3: Examples of classical kernels  $K$  that can be used to define an appropriate density  $\kappa_\rho$  verifying Property 1 in the main paper.

name	support	$K(u)$
Gaussian	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$
Cauchy	$\mathbb{R}$	$\frac{1}{\pi(1+u^2)}$
Laplace	$\mathbb{R}$	$\frac{1}{2} \exp(- u )$
Dirichlet	$\mathbb{R}$	$\frac{\sin^2(u)}{\pi u^2}$
Uniform	$[-1, 1]$	$\frac{1}{2} \mathbb{1}_{ u  \leq 1}$
Triangular	$[-1, 1]$	$(1 -  u ) \mathbb{1}_{ u  \leq 1}$
Epanechnikov	$[-1, 1]$	$\frac{3}{4}(1 - u^2) \mathbb{1}_{ u  \leq 1}$

Another natural choice consists in resorting to a divergence function  $\phi$  such that  $\kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) \propto_{\mathbf{z}} \exp(-\rho^{-1}\phi(\mathbf{z}, \boldsymbol{\theta}))$ . Particular instances of these functions are the family of the Bregman divergences, which are ubiquitous tools in signal processing, machine learning and optimization (see Definition 1 of the main document). Table 4 recalls classical examples of Bregman divergences that can be used to define  $\phi$ . Again, only univariate examples of such potentials are provided but they can be easily extended to the multivariate case.

Note that, reciprocally, tight connections have been already drawn between Bregman divergences and regular exponential family distributions. Indeed, when the function  $\kappa_\rho$  belongs to the exponential family, its associated potential function defined by  $-\log \kappa_\rho$  can

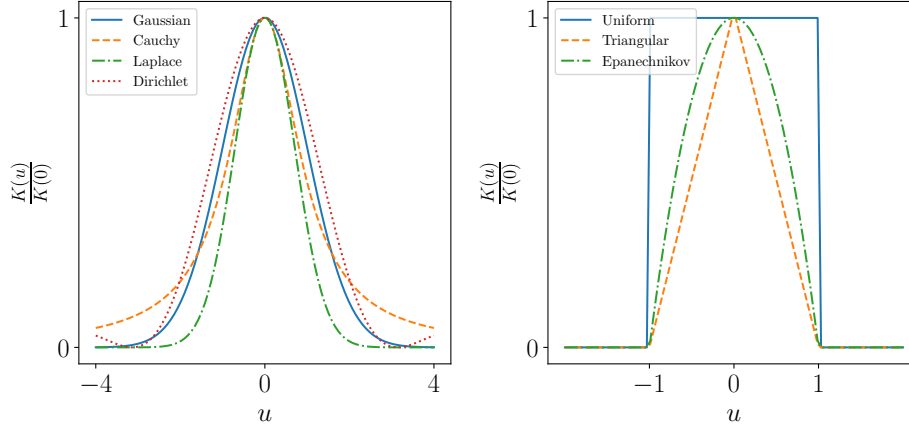


Figure 8: (left) Normalized non compactly-supported kernels; (right) normalized compactly-supported kernels detailed in Table 3.

Table 4: Examples of divergence functions  $\phi$  that can be used to define an appropriate density  $\kappa_\rho$  verifying Property 1 in the main paper.

name	$\Theta$	$\phi(z, \theta)$
Squared loss	$\mathbb{R}$	$(z - \theta)^2$
Absolute loss	$\mathbb{R}$	$ z - \theta $
Logistic loss	$[0, 1]$	$z \log\left(\frac{z}{\theta}\right) + (1 - z) \log\left(\frac{1-z}{1-\theta}\right)$
Itakura-Saito divergence	$\mathbb{R}_+$	$\frac{z}{\theta} - \log\left(\frac{z}{\theta}\right) - 1$
Kullback-Leibler divergence	$[0, 1]$	$z \log\left(\frac{z}{\theta}\right)$

be expressed as a Bregman divergence up to an additional term (Banerjee et al. 2005, Theorem 4)

$$-\log \kappa_\rho(\mathbf{z}, \boldsymbol{\theta}) = d_\psi(\mathbf{z}, \mathbb{E}(\mathbf{z}|\boldsymbol{\theta})) - \log b_\psi(\mathbf{z}). \quad (124)$$

### 3 Illustration for Lipschitz loss functions used in statistical learning

Some of the results of Section 4.3 in the main paper assume that the potential function  $f$  associated to  $\pi$  is Lipschitz. Interestingly, such Lipschitz functions are used in standard

Table 5: Lipschitz loss functions  $f$  used in standard statistical learning problems. Their domain of definition is denoted  $\mathcal{D}_f$  and  $y$  stands for an observation. The notation “reg.” stands for regression.

name	problem	$\mathcal{D}_f$	$f(y; t)$
hinge	SVM	$\{-1, 1\} \times \mathbb{R}$	$\max(0, 1 - yt)$
Huber	robust reg.	$\mathbb{R} \times \mathbb{R}$	$\begin{cases} (y - t)^2 / (2\delta) & \text{if }  y - t  \leq \delta \\  y - t  - \delta/2 & \text{otherwise, where } \delta > 0 \end{cases}$
logistic	logistic reg.	$\{-1, 1\} \times \mathbb{R}$	$\log(1 + \exp(-yt))$
pinball	quantile reg.	$\mathbb{R} \times \mathbb{R}$	$\tau \max(0, t - y) + (1 - \tau) \max(0, y - t), \tau \in (0, 1)$

statistical learning problems to evaluate the discrepancy between observations and model outputs (van de Geer 2016). Table 5 lists some of them along with their definition and associated statistical problems. Note that the absolute loss stands for a particular instance of the pinball loss with  $\tau = 0.5$ . Figure 9 illustrates the form of these losses and associated regularized potentials  $f_\rho$  with  $\rho = 1$  obtained via a Monte Carlo approximation.

Without loss of generality, these problems consider a likelihood function that can be written as in (32) in the main paper with

$$f_j(y_j; \boldsymbol{\theta}) = f(y_j; \mathbf{x}_j^T \boldsymbol{\theta}), \quad (125)$$

where for  $j \in [n]$ ,  $\mathbf{x}_j$  is the feature vector associated with observation  $y_j$ ;  $f$  is one of the loss functions in Table 5 and  $\boldsymbol{\theta} \in \mathbb{R}^d$  is the parameter to infer. Since all the loss functions listed in Table 5 are Lipschitz continuous w.r.t. their second argument  $t$  with Lipschitz constant equal to 1, the potential  $f_j$  in (125) is also Lipschitz with constant  $L_{f_j} = \|\mathbf{x}_j\|_2$ . Motivated by the robustness properties inherited by AXDA, see Section 3.3 in the main paper, we consider the smoothing of the likelihood contribution associated to each observation  $f_j$  with a Gaussian kernel. The results of Corollary 3 in the main paper can then be applied to  $\pi$  defined in (32) in the main paper.

In practice, to illustrate the behavior of the upper bound in Corollary 3 w.r.t. the number of observations, we fixed the dimension  $d$  and considered several values of  $n$  rang-

ing from 1 to  $10^4$ . For each  $n$ , we randomly generated sets of features  $\{\mathbf{x}_j\}_{j \in [n]}$  and we normalized the columns of the matrix  $\mathbf{X}^T = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$  such that each entry is a random number between 0 and 1. The latter operation is classical in machine learning and is also called feature scaling.

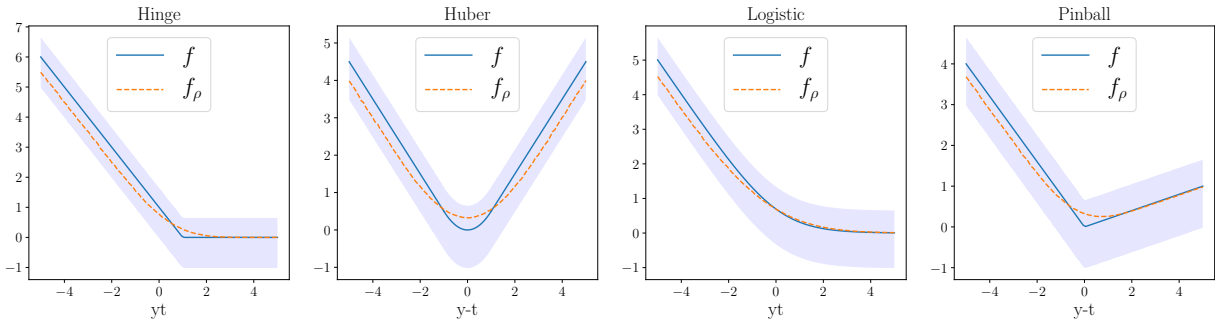


Figure 9: Loss functions of Table 5 along with their associated regularized loss  $f_\rho$  with  $\rho = 1$  estimated with a Monte Carlo approximation. The Huber and pinball losses have been plotted with  $\delta = 1$  and  $\tau = 0.2$ , respectively. The contours of the shaded area correspond to  $f + L_\rho$  and  $f + U_\rho$ .

Figure 10 shows the behavior of the upper bound in Corollary 3 for two values of the dimension  $d = 10$  and  $d = 10^3$ . As expected, the bound becomes less informative for a fixed value of  $\rho$  as the number of likelihood approximations increases with the size of the dataset  $n$ . Nonetheless, the effect of  $n$  on the bound is not highly prohibitive. In both cases  $d = 10$  and  $d = 10^3$ ,  $\rho$  and  $n$  appear to be complementary variables: increasing the value of the latter and decreasing the value of the former by the same factor roughly gives the same bound value. Actually, one can show that the dependence of the bound when  $\rho$  is small is of the order  $\mathcal{O}(n\rho)$  for a fixed dimension  $d$ , see the supplementary material. Obviously, one can limit this dependence on  $n$  by splitting *blocks* of observations in minibatches instead of splitting each observation. This splitting strategy has for instance been considered by Rendell et al. (2020).



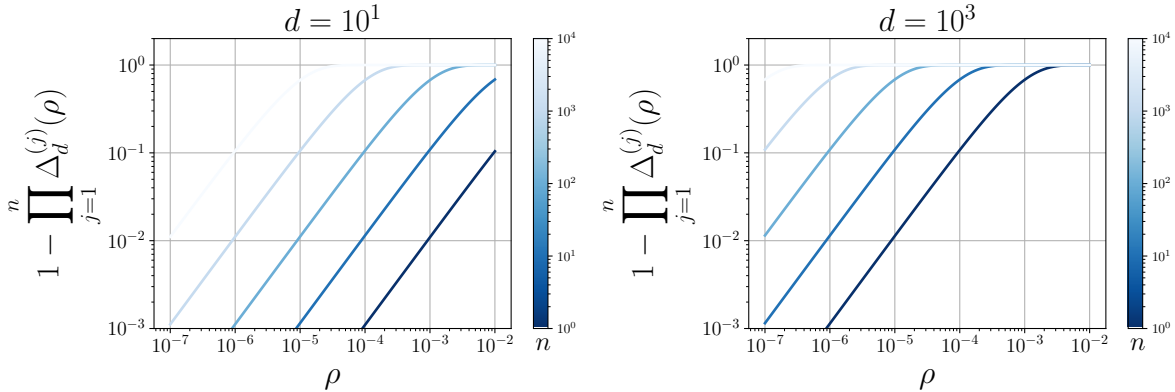


Figure 10: Behavior of the upper bound in Corollary 3 w.r.t.  $\rho$  and  $n$  for several values of the dimension  $d$ . The notation  $\Delta_d^{(j)}(\rho)$  has been defined in Corollary 3.

## 4 Inference details for the image inpainting example

In this section, we detail the steps of the Gibbs sampler used to sample from the posterior distribution  $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$  in Section 5.4 in the main paper.

### 4.1 Sampling the auxiliary vector

The conditional distribution associated to the auxiliary variable  $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2)$  writes

$$\pi_\rho(\mathbf{Z}|\boldsymbol{\theta}) \propto \exp\left(-\tau \sum_{1 \leq i \leq d} \|\mathbf{z}_i\|_2 - \frac{1}{2\rho^2} \|\mathbf{Z} - \mathbf{D}\boldsymbol{\theta}\|_2^2\right). \quad (126)$$

This conditional distribution can be sampled exactly by using data augmentation. Indeed, one can re-write the distribution involving the non-differentiable potential  $\|\cdot\|_2$  as a mixture of normal and gamma distributions (Kyung et al. 2010, Section 3.1). Hence, sampling from (126) can be performed with the following two steps

$$\begin{aligned} &\text{Draw } \frac{1}{\gamma_i} \sim \text{InverseGaussian}\left(\frac{\tau}{\|\mathbf{z}_i\|_2}, \tau^2\right) \forall i \in [d], \text{ if } \|\mathbf{z}_i\|_2 > 0 \\ &\text{Draw } \frac{1}{\gamma_i} \sim \text{InverseGaussian}\left(\frac{3}{2}, \frac{\tau^2}{2}\right) \forall i \in [d], \text{ if } \|\mathbf{z}_i\|_2 = 0 \\ &\text{Draw } \mathbf{z}_{1,i} \sim \mathcal{N}\left(\frac{\gamma_i(D_1\boldsymbol{\theta})_i}{\rho^2 + \gamma_i}, \frac{\rho^2\gamma_i}{\rho^2 + \gamma_i}\right) \forall i \in [d], \\ &\text{Draw } \mathbf{z}_{2,i} \sim \mathcal{N}\left(\frac{\gamma_i(D_2\boldsymbol{\theta})_i}{\rho^2 + \gamma_i}, \frac{\rho^2\gamma_i}{\rho^2 + \gamma_i}\right) \forall i \in [d]. \end{aligned}$$

Note that all these sampling steps can be performed efficiently by “vectorizing” them.

## 4.2 Sampling the parameter of interest

The conditional distribution associated to the image to recover  $\boldsymbol{\theta}$  writes

$$\pi_\rho(\boldsymbol{\theta}|\mathbf{Z}, \mathbf{y}) \propto \exp\left(-\frac{1}{2\rho^2} \|\mathbf{Z} - \mathbf{D}\boldsymbol{\theta}\|_2^2 - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{H}\boldsymbol{\theta}\|_2^2\right). \quad (127)$$

The distribution (127) is a non-degenerate Gaussian distribution  $\mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$  where

$$\boldsymbol{\Sigma}_\theta = \left(\rho^{-2}\mathbf{D}^T\mathbf{D} + \sigma^{-2}\mathbf{H}^T\mathbf{H}\right)^{-1} \quad (128)$$

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta \left(\sigma^{-2}\mathbf{H}^T\mathbf{y} + \rho^{-2}\mathbf{D}^T\mathbf{Z}\right). \quad (129)$$

Sampling from this multivariate distribution can be done efficiently in  $O(d \log d)$  floating point operations by resorting to the two-dimensional discrete Fourier transform. Indeed, under periodic boundary conditions for  $\boldsymbol{\theta}$ , the matrix  $\mathbf{D}^T\mathbf{D}$  is a block circulant matrix and hence diagonalizable in the Fourier domain. On the other hand,  $\mathbf{H}^T\mathbf{H}$  stands for a diagonal matrix with some zeros on the diagonal corresponding to the missing pixels. Since these two matrices cannot be diagonalized in the same domain, we use the auxiliary variable method of [Marnissi et al. \(2018\)](#) to decouple them. Let  $\eta \|\mathbf{H}^T\mathbf{H}\|_S < \sigma^2$  where  $\|\mathbf{M}\|_S$  is the spectral norm of the matrix  $\mathbf{M}$ . Then, we have the following two-step sampling scheme

$$\text{Draw } \mathbf{v} \sim \mathcal{N}\left(\left(\frac{\mathbf{I}_d}{\eta} - \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\right)\boldsymbol{\theta}, \frac{\mathbf{I}_d}{\eta} - \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\right),$$

$$\text{Draw } \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta),$$

where

$$\boldsymbol{\Sigma}_\theta = \left(\frac{\mathbf{I}_d}{\eta} - \frac{\mathbf{D}^T\mathbf{D}}{\rho^2}\right)^{-1},$$

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta \left(\mathbf{v} + \frac{\mathbf{H}^T}{\sigma^2}\mathbf{y} + \frac{\mathbf{D}^T}{\rho^2}\mathbf{Z}\right).$$

## 5 Inference algorithms based on AXDA

Motivated by the issues detailed in Section 2, the good expected properties reviewed in Section 3 as well as the theoretical results shown in Section 4 in the main paper, this section shows that AXDA may allow to derive more efficient and distributed inference algorithms

ranging from simulation to optimization-based methods. To do so, the potential benefits of AXDA in comparison with direct inference from  $\pi$  are presented and discussed. MCMC and VB methods based on AXDA models are detailed to explore the distribution of the parameters to infer. Optimization-based approaches such as ADMM and the EM-algorithm are also derived if maximum likelihood (ML) or maximum a posteriori (MAP) estimates based on the approximate density  $\pi_\rho$  are desired.

From now on, we assume that  $\Theta = \mathbb{R}^d$  and we consider a target density with the general form

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^J \pi_j(\mathbf{A}_j \boldsymbol{\theta}) \propto \prod_{j=1}^J \exp\left(-f_j(\mathbf{A}_j \boldsymbol{\theta})\right). \quad (130)$$

Based on this target density, the augmented density  $\pi_\rho$  is assumed to take the form

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:J}) \propto \prod_{j=1}^J \pi_j(\mathbf{z}_j) \kappa_\rho(\mathbf{z}_j, \mathbf{A}_j \boldsymbol{\theta}). \quad (131)$$

This writing permits to highlight the benefits of using the augmented density  $\pi_\rho$  instead of  $\pi$  for each of the different inference approaches detailed in the sequel.

## 5.1 Monte Carlo sampling from AXDA

A standard way to sample from the joint density  $\pi_\rho$  is to consider a Gibbs sampler as in Algorithm 1. This algorithm can make the inference tractable, simpler and/or faster by targetting  $\pi_\rho$  instead of  $\pi$ . First, by splitting the initial potential  $\sum_j f_j$ ,  $\pi_\rho$  admits simpler and lower-dimensional conditional posterior distributions, each of them possibly defined by a single potential  $f_j$ . Within a Gibbs algorithm, these conditional posteriors yield simpler sampling steps, which may embed efficient dedicated state-of-the-art sampling methods. Second, given the current iterate  $\boldsymbol{\theta}^{[t]}$ , sampling each auxiliary variable  $\mathbf{z}_j^{[t]}$  can be performed in an independent and parallel manner for a faster inference. This is of particular interest in big data settings where datasets are stored on multiple kernels or machines (Rendell et al. 2020). In addition, Vono et al. (2019) experimentally showed that considering AXDA-based models can even improve the convergence properties of classical MCMC methods such as Langevin Monte Carlo by embedding them. A detailed description of additional benefits of AXDA simulation-based methods and their illustration on image

processing and machine learning problems can be found in [Rendell et al. \(2020\)](#) and [Vono et al. \(2019\)](#).

---

**Algorithm 1:** Gibbs sampler

---

**Input:** Functions  $f_j$ , tolerance parameter  $\rho$ , initialization  $\mathbf{z}^{[0]}$  and total nb. of iterations  $T_{\text{MC}}$

```

1 for  $t \leftarrow 1$  to  $T_{\text{MC}}$  do
2   % Drawing the variable of interest  $\boldsymbol{\theta}$ 
3    $\boldsymbol{\theta}^{[t]} \sim \pi_\rho(\boldsymbol{\theta}|\mathbf{z}^{[t-1]}) = \prod_{j=1}^J \kappa_\rho(\mathbf{z}_j^{[t-1]}, \mathbf{A}_j \boldsymbol{\theta})$ ;
4   % Drawing the splitting variables  $\mathbf{z}_j$ 
5   for  $j \leftarrow 1$  to  $J$  do
6     |  $\mathbf{z}_j^{[t]} \sim \pi_\rho(\mathbf{z}_j|\boldsymbol{\theta}^{[t]}) = \pi_j(\mathbf{z}_j) \kappa_\rho(\mathbf{z}_j, \mathbf{A}_j \boldsymbol{\theta}^{[t]})$ ;
7   end
8 end

```

**Output:** Collection of samples  $\{\boldsymbol{\theta}^{[t]}\}_{t=1}^{T_{\text{MC}}}$  asymptotically distributed according to  $\pi_\rho$ .

---

**Example 3.** We consider in this example the penalized logistic regression problem. We assume that  $n$  binary responses  $\mathbf{y} \in \{-1, 1\}^n$  are observed and correspond to conditionally independent Bernoulli random variables with probability of success  $\sigma(\mathbf{x}_j^T \boldsymbol{\theta})$ . The function  $\sigma$  is the sigmoid function,  $\mathbf{x}_j \in \mathbb{R}^d$  stands for the feature vector associated to observation  $y_j$  and  $\boldsymbol{\theta} \in \mathbb{R}^d$  are the unknown regression coefficients to infer. We consider a zero-mean Gaussian prior distribution on  $\boldsymbol{\theta}$  with precision  $2\tau$ , that is  $g(\boldsymbol{\theta}) = \tau \|\boldsymbol{\theta}\|_2^2$ . The target  $\pi$  then stands for the posterior distribution of the unknown regression coefficients  $\boldsymbol{\theta}$

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-g(\boldsymbol{\theta}) - \sum_{j=1}^n \log\left[1 + \exp\left(y_j \mathbf{x}_j^T \boldsymbol{\theta}\right)\right]\right). \quad (132)$$

By denoting  $f_{(n+1)} = g$  and for all  $j \in [n]$  with  $J = n$ ,  $f_j(u) = \log\left[1 + \exp(y_j u)\right]$ , the posterior distribution in (132) has the form (130) with  $J = n + 1$ . Following the work of [Polson et al. \(2013\)](#), one can derive a promising DA scheme from  $\pi$  based on the Poly-Gamma distribution. Hence, a Gibbs sampler can be used to sample from each conditional distribution as detailed by [Polson et al. \(2013\)](#). However, this Gibbs sampler scales poorly in high-dimensional settings as pointed out by [Durmus and Moulines \(2016\)](#). The AXDA

alternative is Algorithm 1 with a quadratic potential and resulting conditional distributions

$$\pi_\rho(z_j|\boldsymbol{\theta}, y_j) \propto \exp\left(-\log\left[1 + \exp(y_j z_j)\right] - \frac{1}{\rho^2}(z_j - \mathbf{x}_j^T \boldsymbol{\theta})^2\right) \quad \forall j \in [n] \quad (133)$$

$$\pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:n}) \propto \exp\left(-\tau \|\boldsymbol{\theta}\|_2^2 - \sum_{j=1}^n \frac{1}{\rho^2}(z_j - \mathbf{x}_j^T \boldsymbol{\theta})^2\right). \quad (134)$$

Thanks to this splitting scheme, the inference is simpler, might be distributed, and sampling from these conditional distributions can be done exactly and efficiently. Indeed, since (133) is univariate and log-concave, one can use adaptive rejection sampling (Gilks and Wild 1992) while sampling the variable of interest  $\boldsymbol{\theta}$  from (134) boils down to high-dimensional Gaussian sampling and efficient methods can be applied.

## 5.2 Variational Bayes inference from AXDA

AXDA can also be a major asset when conducting variational Bayes (VB) inference, providing important benefits such as simplicity and parallelization. VB methods (Bishop and Tipping 2000; Opper and Saad 2001) circumvent the direct sampling from a target density such as  $\pi_\rho$  by defining an approximation of the latter denoted  $\tilde{\pi}_\rho$ . The best approximation is found by minimizing the Kullback-Leibler (KL) divergence between  $\tilde{\pi}_\rho$  and  $\pi_\rho$  restricted to a set of tractable candidates  $\tilde{\pi}_\rho$ . Depending on this set, a lot of VB approximation methods exist in the literature, see Bishop (2006) and Pereyra et al. (2016) for reviews. In this section, we will consider the widely-used mean-field approximation method where the approximate density  $\tilde{\pi}_\rho$  is chosen among the set of conditionally independent (w.r.t.  $\rho$ ) densities, that is  $\tilde{\pi}_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:J}) = \tilde{\pi}_\rho(\boldsymbol{\theta}) \prod_{j=1}^J \tilde{\pi}_\rho(\mathbf{z}_j)$ . Under this constraint, the optimal choice of the VB approximation is given by

$$\log \tilde{\pi}_\rho(\boldsymbol{\theta}) = \sum_{j=1}^J \mathbb{E}_{\tilde{\pi}_\rho(\mathbf{z}_j)} \log \kappa_\rho(\mathbf{z}_j, \mathbf{A}_j \boldsymbol{\theta}) \quad (135)$$

$$\log \tilde{\pi}_\rho(\mathbf{z}_j) = -f_j(\mathbf{z}_j) + \mathbb{E}_{\tilde{\pi}_\rho(\boldsymbol{\theta})} \log \kappa_\rho(\mathbf{z}_j, \mathbf{A}_j \boldsymbol{\theta}). \quad (136)$$

The above VB-marginals require to compute expectations under each marginal distribution which are often functions of moments under each marginal. Similarly to what has been encountered for Gibbs sampling in the previous section, deriving a VB approach based on  $\pi_\rho$ , instead of  $\pi$ , yields important benefits for parallel and possibly easier computations.

Indeed, the VB-marginal in (136) shows again that each potential  $f_j$  contributes independently given  $\boldsymbol{\theta}$ . The updates of expectations under (136) are thereby simplified since (i) the VB-marginals (136) are simpler than those obtained from a mean-field approximation of  $\pi$  and (ii) the moments under the latter can be computed in parallel or distributed.

**Example 3 (continued).** Jaakkola and Jordan (2000) considered a local VB algorithm for the penalized logistic regression problem. Instead of using a local VB approach and finding bounds on each individual function  $f_j$ , the use of  $\pi_\rho$  instead of  $\pi$  permits to consider directly a global VB approach such as the mean-field approximation. In addition, similarly to Algorithm 1, the updates of (136) and the associated expectations can be computed in parallel and efficiently by using state-of-the-art existing methods. For instance, the expectations  $\mathbb{E}_{\tilde{\pi}_\rho(\mathbf{z}_j)}$  can be approximated efficiently using rejection sampling.

### 5.3 Optimizing AXDA meets quadratic penalty methods

Computing the MAP or ML estimate under the AXDA model (131) boils down to solve the optimization problem

$$\min_{\boldsymbol{\theta}, \mathbf{z}_{1:J}} \sum_{j=1}^J f_j(\mathbf{z}_j) - \log \kappa_\rho(\mathbf{z}_j, \mathbf{A}_j \boldsymbol{\theta}). \quad (137)$$

If  $\kappa_\rho$  stands for a Gaussian kernel, the problem (137) can be viewed as a quadratically penalized formulation of the initial problem  $\min_{\boldsymbol{\theta}} \sum_{j=1}^J f_j(\boldsymbol{\theta})$ , see Nocedal and Wright (2006, Section 17.1). As expected, the solution of (137) stands for an approximate solution w.r.t. the initial optimization problem. The associated algorithm is depicted in Algorithm 2. Regarding this algorithm, one can clearly see the benefit of using a variable splitting approach as in AXDA: the initial potential is split into  $J$  individual potentials with no operator acting on  $\boldsymbol{\theta}$ . Therefore, the corresponding minimization problems are simpler (e.g., associated proximity operators may become available) and can be handled in parallel.

We eventually point out that the benefits of Algorithm 2 highlighted previously are also shared with the ADMM (Boyd et al. 2011). Instead of solving the approximate optimization problem (137) which encodes the splitting operation with a quadratic regularization term, the latter builds on Lagrangian duality in order to provide an exact solution to the initial minimization problem  $\min f_j$ .

---

**Algorithm 2:** Quadratic penalty minimization

---

**Input:** Functions  $f_j$ , penalty parameter  $\rho$ ,  $t \leftarrow 0$  and  $\mathbf{z}_{1:J}^{[0]}$

```
1 while stopping criterion not satisfied do
2   % Minimization w.r.t.  $\boldsymbol{\theta}$ 
3    $\boldsymbol{\theta}^{[t]} \in \arg \min_{\boldsymbol{\theta}} \sum_{j=1}^J \frac{1}{2\rho^2} \|\mathbf{A}_j \boldsymbol{\theta} - \mathbf{z}_j^{[t-1]}\|_2^2$ ;
4   for  $j \leftarrow 1$  to  $J$  do
5     % Minimization w.r.t.  $\mathbf{z}_j$ 
6      $\mathbf{z}_j^{[t]} \in \arg \min_{\mathbf{z}_j} f_j(\mathbf{z}_j) + \frac{1}{2\rho^2} \|\mathbf{A}_j \boldsymbol{\theta}^{[t]} - \mathbf{z}_j\|_2^2$ ;
7   end
8   % Updating iterations counter
9    $t \leftarrow t + 1$ ;
10 end
```

**Output:** MAP or ML estimate depending on the considered problem.

---

**Example 3 (continued).** Computing directly the MAP estimate with classical forward-backward algorithms (e.g., the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle 2009)) associated to  $\pi$  is challenging if the observations are distributed among multiple nodes. In addition, proximity operators associated to  $f_j$  are generally not available in closed-form because of the operators  $\mathbf{A}_j$ . Algorithm 2 permits to tackle these issues by splitting the initial objective function. In particular, the minimization w.r.t.  $\mathbf{z}_j$  for  $j \in [J]$  corresponds to an unidimensional  $l_2$  regularized logistic regression problem that can be dealt with gradient-based methods with few iterations. The minimization w.r.t.  $\boldsymbol{\theta}$  boils down to the solving of a linear system where efficient solvers can be applied. Note that such a splitting scheme avoids the use of stochastic gradient methods.

## 5.4 Expectation-maximization for AXDA

An EM algorithm under the augmented density  $\pi_\rho(\boldsymbol{\theta}, \mathbf{z})$  will target the MAP or ML estimator, see Algorithm 3. If the expectations in the E-step cannot be evaluated, one can use a Monte Carlo approximation to approximate them (Wei and Tanner 1990). The benefits of using the augmented density  $\pi_\rho$  instead of  $\pi$  are threefolds. Firstly, as pointed out in Section 2 in the main paper, exact DA schemes based on  $\pi$  cannot be derived in general cases and corresponding EM algorithms cannot be implemented. Instead, considering  $\pi_\rho$  gives a quite systematic way of introducing latent variables in the original statistical model. Secondly, the expectations involved in the E-step of Algorithm 3 can be simpler to derive

than the expectation under  $\pi$ . Indeed, the latter involves the whole potential  $\sum_j f_j$  while the former involves regularized parts  $f_j - \log \kappa_\rho$  of this potential separately. Finally, conditionally on  $\boldsymbol{\theta}^{[t]}$ , the random variables  $\mathbf{z}_j$  are independent. Thus, each expectation involved in the E-step can be computed in parallel.

---

**Algorithm 3:** EM

---

**Input:** Functions  $f_j$ , penalty parameter  $\rho$ ,  $t \leftarrow 0$  and  $\boldsymbol{\theta}^{[0]}$

1 **while** *stopping criterion not satisfied* **do**

2     % E-step

3     Define  $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}) = \sum_{j=1}^J \mathbb{E}_{\pi_\rho(\mathbf{z}_j|\boldsymbol{\theta}^{[t]})} \left( -f_j(\mathbf{z}_j) + \log \kappa_\rho(\mathbf{z}_j, \mathbf{A}_j \boldsymbol{\theta}) \right);$

4     % M-step

5     ; Compute  $\boldsymbol{\theta}^{[t+1]} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]});$

6     % Updating iterations counter

7      $t \leftarrow t + 1$  ;

8 **end**

**Output:** MAP or ML estimate depending on the considered problem.

---

**Example 3 (continued).** Again, following the work of Polson et al. (2013), if the potential  $g$  of the prior distribution is quadratic or corresponds to a sparsity-promoting  $\ell_p$ -penalization ( $0 < p \leq 1$ ), a simple EM-algorithm can be derived as detailed by Scott and Sun (2013). However, although this EM algorithm can be generalized to an online version, it does not scale to distributed and high-dimensional problems. On the other hand, the E-step of Algorithm 3 can be processed in parallel by computing the  $J$  expectations on individual nodes: thanks to the AXDA approach, the algorithm is therefore suited to distributed and high-dimensional scenarios.



## References

- Aguirregabiria, J. M., Hernández, A., and Rivas, M. (2002), “ $\delta$ -function converging sequences,” *American Journal of Physics*, 70, 180–185.
- Albert, J. H. and Chib, S. (1993), “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88, 669–679.
- Ambrosio, L., Gigli, N., and Savaré, G. (2008), *Gradient flows in metric spaces and in the space of probability measures*, Lectures in Mathematics. ETH Zürich, Birkhäuser Verlag, 2nd ed.
- Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005), “Clustering with Bregman Divergences,” *Journal of Machine Learning Research*, 6, 1705–1749.
- Baragatti, M., Grimaud, A., and Pommeret, D. (2013), “Likelihood-free parallel tempering,” *Statistics and Computing*, 23, 535–549.
- Barbos, A.-C., Caron, F., Giovannelli, J.-F., and Doucet, A. (2017), “Clone MCMC: parallel high-dimensional Gaussian Gibbs sampling,” in *Advances in Neural Information Processing Systems*, pp. 5020–5028.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), “Approximate Bayesian Computation in Population Genetics,” *Genetics*, 162, 2025–2035.
- Beck, A. and Teboulle, M. (2009), “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems,” *SIAM Journal on Imaging Sciences*, 2, 183–202.
- Ben-Tal, A., Margalit, T., and Nemirovski, A. (2001), “The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography,” *SIAM Journal on Optimization*, 12, 79–108.
- Besag, J. and Green, P. J. (1993), “Spatial Statistics and Bayesian Computation,” *Journal of the Royal Statistical Society, Series B*, 55, 25–37.
- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer-Verlag.

- Bishop, C. M. and Tipping, M. E. (2000), “Variational Relevance Vector Machines,” in *Conference on Uncertainty in Artificial Intelligence*.
- Bolte, J., Sabach, S., and Teboulle, M. (2014), “Proximal alternating linearized minimization for nonconvex and nonsmooth problems,” *Mathematical Programming*, 146, 459–494.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Foundations and Trends in Machine Learning*, 3, 1–122.
- Bredies, K., Kunisch, K., and Pock, T. (2010), “Total Generalized Variation,” *SIAM Journal on Imaging Sciences*, 3, 492–526.
- Canny, J. (2004), “GaP: A Factor Model for Discrete Data,” in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 122–129.
- Chambolle, A. (2004), “An Algorithm for Total Variation Minimization and Applications,” *Journal of Mathematical Imaging and Vision*, 20, 89–97.
- Chambolle, A., Novaga, M., Cremers, D., and Pock, T. (2010), “An introduction to total variation for image analysis,” in *Theoretical Foundations and Numerical Methods for Sparse Recovery*, De Gruyter.
- Choi, H. M. and Hobert, J. P. (2013), “The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic,” *Electronic Journal of Statistics*, 7, 2054–2064.
- Combettes, P. L. and Pesquet, J.-C. (2011), “Proximal Splitting Methods in Signal Processing,” in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, eds. Bauschke, H. H., Burachik, R., Combettes, P., Elser, V., Luke, D., and Wolkowicz, H., Springer, pp. 185–212.
- Damien, P., Wakefield, J., and Walker, S. (1999), “Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables,” *Journal of the Royal Statistical Society, Series B*, 61, 331–344.

- Dang, Q. A. and Ehrhardt, M. (2012), “On Dirac delta sequences and their generating functions,” *Applied Mathematics Letters*, 25, 2385 – 2390.
- Del Moral, P., Doucet, A., and Jasra, A. (2012), “An adaptive sequential Monte Carlo method for approximate Bayesian computation,” *Statistics and Computing*, 22, 1009–1020.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Dharmadhikari, S. and Joag-Dev, K. (1988), *Unimodality, Convexity, and Applications*, New York: Academic Press.
- Dobson, A. J. and Barnett, A. G. (2008), *An Introduction to Generalized Linear Models*, Texts in Statistical Science, Boca Raton, FL: Chapman & Hall/CRC Press, 3rd ed.
- Doucet, A., Godsill, S. J., and Robert, C. P. (2002), “Marginal maximum a posteriori estimation using Markov chain Monte Carlo,” *Statistics and Computing*, 12, 77–84.
- Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987), “Hybrid Monte Carlo,” *Physics Letters B*, 195, 216–222.
- Dümbgen, L. and Rufibach, K. (2009), “Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency,” *Bernoulli*, 15, 40–68.
- Durmus, A. and Moulines, E. (2016), “High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm,” [online]. Technical report. Available at <https://arxiv.org/abs/1605.01559/>.
- (2017), “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm,” *The Annals of Applied Probability*, 27, 1551–1587.
- Durmus, A., Moulines, E., and Pereyra, M. (2018), “Efficient Bayesian Computation by Proximal Markov chain Monte Carlo: When Langevin Meets Moreau,” *SIAM Journal on Imaging Sciences*, 11, 473–506.

- Edwards, R. G. and Sokal, A. D. (1988), “Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm,” *Physical Review D*, 38, 2009–2012.
- Fearnhead, P. and Prangle, D. (2012), “Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation,” *Journal of the Royal Statistical Society, Series B*, 74, 419–474.
- Fellows, M., Mahajan, A., Rudner, T. G. J., and Whiteson, S. (2019), “VIREL: A Variational Inference Framework for Reinforcement Learning,” in *Advances in Neural Information Processing Systems*, pp. 7120–7134.
- Folland, G. (1999), *Real Analysis: Modern Techniques and their Applications*, New York: Wiley, 2nd ed.
- Geman, D. and Reynolds, G. (1992), “Constrained restoration and the recovery of discontinuities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 367–383.
- Geman, D. and Yang, C. (1995), “Nonlinear image recovery with half-quadratic regularization,” *IEEE Transactions on Image Processing*, 4, 932–946.
- Gilks, W. R. and Wild, P. (1992), “Adaptive Rejection Sampling for Gibbs Sampling,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41, 337–348.
- Giovannelli, J. F. (2008), “Unsupervised Bayesian Convex Deconvolution Based on a Field With an Explicit Partition Function,” *IEEE Transactions on Image Processing*, 17, 16–26.
- Gradshteyn, I. S. and Ryzhik, I. M. (2015), *Table of Integrals, Series, and Products*, London: Academic Press, 8th ed.
- Hartley, H. O. (1958), “Maximum Likelihood Estimation from Incomplete Data,” *Biometrics*, 14, 174–194.
- Higdon, D. (2007), *A Primer on Space-Time Modeling from a Bayesian Perspective*, Chapman & Hall/CRC, pp. 217–279.

- Higdon, D. M. (1998), “Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications,” *Journal of the American Statistical Association*, 93, 585–595.
- Holmes, C. C. and Mallick, B. K. (2003), “Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines,” *Journal of the American Statistical Association*, 98, 352–368.
- Hurn, M. (1997), “Difficulties in the use of auxiliary variables in Markov chain Monte Carlo methods,” *Statistics and Computing*, 7, 35–44.
- Jaakkola, T. S. and Jordan, M. I. (2000), “Bayesian parameter estimation via variational methods,” *Statistics and Computing*, 10, 25–37.
- Krichene, W., Bayen, A., and Bartlett, P. L. (2015), “Accelerated Mirror Descent in Continuous and Discrete Time,” in *Advances in Neural Information Processing Systems*, pp. 2845–2853.
- Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), “Penalized regression, standard errors, and Bayesian lassos,” *Bayesian Analysis*, 5, 369–411.
- Liechty, M. W., Liechty, J. C., and Müller, P. (2009), “The Shadow Prior,” *Journal of Computational and Graphical Statistics*, 18, 368–383.
- Marnissi, Y., Chouzenoux, E., Benazza-Benyahia, A., and Pesquet, J.-C. (2018), “An Auxiliary Variable Method for Markov Chain Monte Carlo Algorithms in High Dimension,” *Entropy*, 20.
- Meng, X.-L. and van Dyk, D. (1997), “The EM Algorithm – an Old Folk-song Sung to a Fast New Tune,” *Journal of the Royal Statistical Society, Series B*, 59, 511–567.
- Nesterov, Y. and Spokoiny, V. (2017), “Random Gradient-Free Minimization of Convex Functions,” *Foundations of Computational Mathematics*, 17, 527–566.
- Nocedal, J. and Wright, S. J. (2006), *Numerical Optimization*, Springer, 2nd ed.
- Opper, M. and Saad, D. (2001), *Advanced Mean Field Methods: Theory and Practice*, MIT Press.

- Papandreou, G. and Yuille, A. L. (2010), “Gaussian sampling by local perturbations,” in *Advances in Neural Information Processing Systems*, pp. 1858–1866.
- Park, T. and Casella, G. (2008), “The Bayesian Lasso,” *Journal of the American Statistical Association*, 103, 681–686.
- Pereyra, M. (2016), “Proximal Markov chain Monte Carlo algorithms,” *Statistics and Computing*, 26, 745–760.
- Pereyra, M., Schniter, P., Chouzenoux, E., Pesquet, J.-C., Tourneret, J.-Y., Hero, A., and McLaughlin, S. (2016), “A Survey of Stochastic Simulation and Optimization Methods in Signal Processing,” *IEEE Journal of Selected Topics in Signal Processing*, 10, 224–241.
- Polson, N. G. (1996), “Convergence of Markov chain Monte Carlo algorithms,” in *Bayesian Statistics 5*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., pp. 297–321.
- Polson, N. G., Scott, J. G., and Windle, J. (2013), “Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables,” *Journal of the American Statistical Association*, 108, 1339–1349.
- Rendell, L. J., Johansen, A. M., Lee, A., and Whiteley, N. (2020), “Global consensus Monte Carlo,” *Journal of Computational and Graphical Statistics*.
- Robert, C. P. (2001), *The Bayesian Choice: from decision-theoretic foundations to computational implementation*, New York: Springer, 2nd ed.
- Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Berlin: Springer, 2nd ed.
- Salim, A., Koralev, D., and Richtarik, P. (2019), “Stochastic Proximal Langevin Algorithm: Potential Splitting and Nonasymptotic Rates,” in *Advances in Neural Information Processing Systems*, pp. 6649–6661.
- Scheffé, H. (1947), “A useful convergence theorem for probability distributions,” *The Annals of Mathematical Statistics*, 18, 434–438.

- Scott, J. G. and Sun, L. (2013), “Expectation-maximization for logistic regression,” [online]. Technical report. Available at <https://arxiv.org/abs/1306.0040/>.
- Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016), “Bayes and Big Data: The Consensus Monte Carlo Algorithm,” *International Journal of Management Science and Engineering Management*, 11, 78–88.
- She, Y. and Owen, A. B. (2011), “Outlier Detection Using Nonconvex Penalized Regression,” *Journal of the American Statistical Association*, 106, 626–639.
- Sisson, S., Fan, Y., and Beaumont, M. (eds.) (2018a), *Handbook of Approximate Bayesian Computation*, Chapman and Hall/CRC Press, 1st ed.
- Sisson, S. A., Fan, Y., and Beaumont, M. A. (2018b), “Overview of Approximate Bayesian Computation,” in *Handbook of Approximate Bayesian Computation*, ed. Author, Chapman and Hall/CRC Press, 1st ed., pp. 3–54.
- Swendsen, R. H. and Wang, J.-S. (1987), “Nonuniversal critical dynamics in Monte Carlo simulations,” *Physical Review Letters*, 58, 86–88.
- Tanner, M. A. and Wong, W. H. (1987), “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82, 528–540.
- (2010), “From EM to Data Augmentation: The Emergence of MCMC Bayesian Computation in the 1980s,” *Statistical Science*, 25, 506–516.
- van de Geer, S. (2016), *Estimation and Testing Under Sparsity*, Lecture Notes in Mathematics 2159, Springer, 1st ed.
- van Dyk, D. A. and Meng, X.-L. (2001), “The Art of Data Augmentation,” *Journal of Computational and Graphical Statistics*, 10, 1–50.
- Vono, M., Dobigeon, N., and Chainais, P. (2019), “Split-and-augmented Gibbs sampler - Application to large-scale inference problems,” *IEEE Transactions on Signal Processing*, 67, 1648–1661.
- Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman & Hall/CRC.

- Wang, C. and Blei, D. M. (2018), “A General Method for Robust Bayesian Modeling,” *Bayesian Analysis*, 13, 1163–1191.
- Wang, X. and Dunson, D. B. (2013), “Parallelizing MCMC via Weierstrass Sampler,” [online]. Technical report. Available at <https://arxiv.org/abs/1312.4605/>.
- Wang, Y., Yang, J., Yin, W., and Zhang, Y. (2008), “A New Alternating Minimization Algorithm for Total Variation Image Reconstruction,” *SIAM Journal on Imaging Sciences*, 1, 248–272.
- Wei, G. C. G. and Tanner, M. A. (1990), “A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithms,” *Journal of the American Statistical Association*, 85, 699–704.
- Wilkinson, R. (2013), “Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error,” *Statistical applications in genetics and molecular biology*, 12, 1–13.