THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Institut National Polytechnique de Toulouse (Toulouse INP)
Discipline ou spécialité :
Signal, Image, Acoustique et Optimisation

Présentée et soutenue par :

M. MAXIME VONO

le mercredi 7 octobre 2020

Titre :

Asymptotically exact data augmentation – Models and Monte Carlo sampling with applications to Bayesian inference

---

Devant le jury d'examen :

Rapporteurs	
Mme FLORENCE FORBES	Directrice de Recherche INRIA, Université Grenoble Alpes
M. JALAL FADILI	Professeur des Universités, ENSI Caen
Examinateurs	
Mme ÉMILIE CHOUZENOUX	Chargée de Recherche INRIA, CentraleSupelec
M. JEAN-MICHEL MARIN	Professeur des Universités, Université de Montpellier
M. ÉRIC MOULINES	Professeur des Universités, École Polytechnique
Directeurs de thèse	
M. NICOLAS DOBIGEON	Professeur des Universités, INP-ENSEEIHT
M. PIERRE CHAINAIS	Professeur des Universités, Centrale Lille

Thèse préparée dans le Laboratoire :
Institut de Recherche en Informatique de Toulouse (IRIT)
École doctorale :
Mathématiques, Informatique, Télécommunications de Toulouse (MITT)

À mes parents.

Remerciements

Bien qu'étant un projet individuel, de nombreuses personnes ont contribué à la réussite de cette thèse. J'espère que ces quelques mots sauront leur faire honneur et les remercier comme il se doit.

Pierre, c'est avec toi que tout a commencé. J'ai reçu un mail de ta part pendant mon stage de fin d'études pour un projet de thèse en collaboration avec Toulouse. Malgré quelques péripéties, j'ai accepté et l'aventure a commencé le lundi 2 octobre 2017. Un grand merci pour m'avoir fait confiance, encadré, fait grandir et toujours soutenu dans les moments difficiles. Ces trois années ont été un réel plaisir.

Nicolas, j'ai été très chanceux et honoré de travailler avec toi. J'aimerais te remercier sincèrement pour ton encadrement, ta confiance, tes critiques toujours constructives et ton envie d'aller toujours plus loin. C'est avec émotion que je quitte ton équipe. Un petit clin d'oeil également pour m'avoir placé (stratégiquement ?) dans le bureau F418 !

Je souhaiterais ensuite remercier les membres de mon jury de thèse pour leur investissement et leurs critiques constructives. Un grand merci à Florence Forbes et Jalal Fadili pour avoir accepté de rapporter cette thèse. Je remercie également Émilie Chouzenoux, Jean-Michel Marin et Éric Moulines pour avoir accepté d'examiner cette thèse. J'ai été honoré de vous avoir au sein de mon jury.

Durant cette thèse, j'ai également eu la chance de passer quelques mois au Department of Statistics de l'Université d'Oxford. Un grand merci à Arnaud Doucet qui m'a accueilli et à Daniel Paulin avec qui j'ai étroitement collaboré. C'était un réel plaisir de travailler avec vous.

L'incroyable environnement dans lequel j'ai évolué durant ces trois années a énormément joué dans la réussite de cette thèse. Merci à Cédric, Emmanuel, Thomas et Marie. Merci aux doctorant(e)s et post-doctorant(e)s passés et présents : Louis, Olivier, Alberto, Yanna, Vinicius, Jessica, Dylan, Serdar, Adrien, Étienne, Baha, Mouna, Claire, Camille, Asma, Pierre-Hugo, Cassio, Sixin, Paul, Dana. J'aimerais également grandement remercier Annabelle et son travail méticuleux.

Merci également aux collègues lillois que j'ai connus pendant mes années à Centrale et avec qui j'ai passé plusieurs semaines durant ma première année de thèse : Patrick, Rémi, Julien, Clément, John, Ayoub, Quentin, Phuong, Victor, François. Un grand merci à Guillaume.

Avant de conclure, j'aimerais avoir un remerciement particulier pour mes parents et ma soeur qui m'ont toujours soutenu. C'est avec émotion que j'ai soutenu cette thèse devant vous.

Enfin, merci Clémentine, pour tout.

Résumé

De nombreuses tâches d'apprentissage statistique et de traitement du signal/de l'image peuvent être formulées comme des problèmes d'inférence statistique. L'objet à estimer est généralement défini comme la solution d'un problème d'optimisation variationnelle ou stochastique. En particulier, dans un cadre bayésien, cette solution est définie comme le minimiseur d'une fonction de coût, appelée fonction de perte a posteriori. Dans le cas simple où cette fonction est choisie comme quadratique, l'estimateur bayésien est connu pour être la moyenne a posteriori qui minimise l'erreur quadratique moyenne et qui est définie comme une intégrale par rapport à la distribution a posteriori. Dans la plupart des contextes applicatifs du monde réel, le calcul de telles intégrales n'est pas simple. Une alternative consiste à utiliser l'intégration de Monte Carlo, qui se résume à approximer toute espérance par une moyenne empirique impliquant des échantillons générés selon la distribution cible. Cette intégration dite de Monte Carlo nécessite la disponibilité de schémas algorithmiques efficaces capables de générer des échantillons à partir d'une distribution a posteriori souhaitée. Une vaste littérature consacrée à la génération de variables aléatoires a proposé divers algorithmes de Monte Carlo. Par exemple, les méthodes de Monte Carlo par chaîne de Markov (MCMC), dont les exemples particuliers sont le célèbre échantillonneur de Gibbs et l'algorithme de Metropolis-Hastings, définissent une large classe d'algorithmes qui permettent de générer une chaîne de Markov avec la distribution stationnaire souhaitée. Malgré leur simplicité et leur caractère générique en apparence, les algorithmes MCMC classiques peuvent se révéler inefficaces pour les problèmes en grande dimension, distribués et/ou très structurés.

L'objectif principal de cette thèse consiste à introduire de nouveaux modèles et approches MCMC pour pallier ces problèmes. La complexité de la distribution a posteriori est abordée en proposant une classe de modèles augmentés approchés mais asymptotiquement exacts (AXDA). Ensuite, un échantillonneur de Gibbs ciblant une distribution a posteriori approchée construite dans le cadre AXDA est proposé et ses avantages sont illustrés sur des problèmes difficiles de traitement du signal, de traitement d'image et d'apprentissage statistique. Une étude théorique détaillée du taux de convergence associé à cet échantillonneur est également menée et révèle des dépendances explicites à la dimension, au conditionnement du potentiel de la loi a posteriori et à la précision prescrite. Dans ce travail, nous prêtons également attention à la faisabilité des étapes d'échantillonnage impliquées dans l'échantillonneur de Gibbs proposé. Comme l'une de ces étapes nécessite d'échantillonner selon une distribution gaussienne en grande dimension, nous passons en revue et unifions les approches existantes en introduisant un cadre qui s'interprète comme la contrepartie stochastique du célèbre algorithme du point proximal. Ce lien fort entre la simulation et l'optimisation n'est pas isolé dans cette thèse. En effet, nous montrons également que les échantillonneurs de Gibbs proposés partagent des liens étroits avec les méthodes de pénalité quadratique et que le cadre AXDA génère une classe de fonctions d'enveloppe liées à celle de Moreau.

Mots-clés : Optimisation, algorithmes de Monte Carlo, statistiques, traitement du signal

Abstract

Many statistical learning and signal/image processing tasks can be formulated as statistical inference problems. A typical example are recommendation systems that are based on the completion of a partially observed user/object matrix, which can be achieved by the joint estimation of latent factors and activation coefficients. More formally, the object to be estimated is generally defined as the solution of a variational or stochastic optimization problem. In particular, in a Bayesian framework, this solution is defined as the minimizer of a cost function, called the posterior expected loss. In the simple case where this function is chosen as quadratic, the Bayesian estimator is known to be the posterior mean that minimizes the mean square error and is defined as an integral with respect to the posterior distribution. In most real-world applicative contexts, the computation of such integrals is not simple. An alternative is to use Monte Carlo integration, which boils down to approximating any expectation by an empirical mean involving samples generated according to this target distribution. This Monte Carlo integration requires the availability of efficient algorithmic schemes capable of generating samples from a desired distribution. An extensive literature devoted to the generation of random variables has proposed various Monte Carlo algorithms. For example, Markov chain Monte Carlo methods (MCMC), of which the famous Gibbs sampler and the Metropolis-Hastings algorithm are particular examples, define a broad class of algorithms that generate a Markov chain with the desired stationary distribution. Despite their simplicity and seemingly generic nature, conventional MCMC algorithms may be ineffective for large, distributed and/or highly structured problems.

The main objective of this thesis is to introduce new MCMC models and approaches to overcome these problems. The intractability of the posterior distribution is addressed by proposing a class of approximate but asymptotically exact augmented models (AXDA). Next, a so-called split Gibbs sampler (SGS) targeting the approximate posterior distribution built via the AXDA framework is proposed and its advantages are illustrated on difficult signal processing, image processing and statistical learning problems. A detailed theoretical study of the convergence rate of SGS is also conducted and reveals explicit dependencies with respect to the dimension of the problem, the conditioning of the potential function of the posterior distribution and the prescribed tolerance. In this work, we also pay attention to the feasibility of the sampling steps involved in the proposed Gibbs sampler. Since one of these steps requires sampling according to a high-dimensional Gaussian distribution, we review and unify existing approaches by introducing a framework that can be interpreted as the stochastic counterpart of the celebrated proximal point algorithm. This strong link between simulation and optimization is not isolated in this thesis. Indeed, we also show that the proposed Gibbs sampler shares close links with quadratic penalty methods and that the AXDA framework generates a class of envelope functions linked to Moreau's one.

Keywords : Optimization, Monte Carlo algorithms, statistics, signal processing

Contents

Notations

Some specific sets

$\mathbb{R}$	Real numbers
$\mathbb{R}^n$	Real n-vectors ($n \times 1$ matrices)
$\mathbb{R}^{m \times n}$	Real $m \times n$ matrices
$\mathcal{B}(\mathbb{R}^d)$	Borel σ-field of $\mathbb{R}^d$
$\mathbb{M}(\mathbb{R}^d)$	Set of all Borel measurable functions f on $\mathbb{R}^d$
L^1	Set of all Borel measurable functions f on $\mathbb{R}^d$ such that their absolute value is Lebesgue integrable
$[n]$	Set of all postive integers between 1 and n
$[a, b]$	Real interval containing a and b
$[a, b)$	Real interval containing a but excluding b

Norms and distances

$\|\mathbf{a}\|$	Euclidean norm of the vector $\mathbf{a}$
$\|\mathbf{a}\|_1$	ℓ_1 norm of the vector $\mathbf{a}$
$\|f\|_\infty$	Infinity norm of the function f
$\|\mu - \nu\|_{\mathrm{TV}}$	Total variation distance between the measures μ and ν
$W_p(\mu, \nu)$	p-Wasserstein distance between the measures μ and ν

Probability

$\mathcal{U}(\cdot; [a, b])$	Uniform distribution over the interval $[a, b]$
$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian (also called normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{B}(\cdot; p)$	Bernoulli distribution with probability of success p
$\mathcal{P}(\cdot; \lambda)$	Poisson distribution with intensity λ
$\mathcal{IG}(\cdot; a, b)$	Inverse-gamma distribution with shape a and scale b
$\mathrm{Gumbel}(\cdot; \mu, \sigma)$	Gumbel distribution with location μ and scale σ
$\mathrm{InverseGaussian}(\cdot; \mu, \lambda)$	Inverse Gaussian distribution with location μ and shape λ
$\mathbb{P}(X \in \mathcal{A})$	Probability that X belongs to $\mathcal{A}$
$\mathbb{E}_\pi(X)$	Expected value of X under π

Indexing

a_i	i-th element of the vector $\mathbf{a}$
A_{ij}	Element (i, j) of the matrix $\mathbf{A}$
$\mathbf{a}_{i:j}$	Vector $[\mathbf{a}_i, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_j]^T$ built by stacking $j - i + 1$ vectors $(\mathbf{a}_k; k \in \{i, i+1, \ldots, j\})$

Linear algebra

$\mathbf{A}^T$	Transpose of the matrix $\mathbf{A}$
$\mathrm{Trace}(\mathbf{A})$	Trace of the matrix $\mathbf{A}$
$\det(\mathbf{A})$	Determinant of the matrix $\mathbf{A}$

Introduction

Statistical inference and Monte Carlo integration

Numerous tasks can be formulated as statistical inference problems which aim to draw insightful conclusions from observations of a random phenomenon about quantities that are not observed. In machine learning, archetypal examples are recommendation systems that are based on the completion of a partially observed user/object matrix, which can be conducted via the joint estimation of latent (i.e., unobserved) factors and activation coefficients (Candes and Plan, 2010; Melville and Sindhwani, 2010). Similarly, ubiquitous signal/image processing tasks are usually formulated as the estimation of latent objects or features, whether for low-level processing (e.g., denoising, deconvolution, restoration) or for high-level analysis (e.g., classification, segmentation, feature extraction) (Idier, 2008). More formally, the object θ to be inferred is usually defined as the solution of a variational or stochastic optimization problem. In particular within a Bayesian framework (Robert, 2001), the estimated solution $\hat{\theta}$ is defined as the minimizer of a cost function, referred to as the posterior loss and defined as

$$\hat{\theta} \in \arg\min_{\delta} \mathbb{E}\left[L(\delta,\theta)|y\right], \quad \mathbb{E}\left[L(\delta,\theta)|y\right] = \int L(\delta,\theta)\pi(\theta|y)\mathrm{d}\theta, \quad (0.1)$$

where y denotes the set of available data modeled as the realization of a random variable Y fully characterized by the likelihood function $\pi(y|\theta)$, $\pi(\theta|y)$ is the posterior distribution related to the likelihood function $\pi(y|\theta)$ and prior distribution $\pi(\theta)$ thanks to the Bayes formula and $L(\cdot,\cdot)$ is a given loss function. When this function is chosen as quadratic, i.e., $L(\delta,\theta) \coloneqq \|\delta-\theta\|^2$, the Bayesian estimator $\hat{\theta}$ is known to be the posterior mean $\hat{\theta}_{\mathrm{MMSE}} = \mathbb{E}[\theta|y]$, expressed via an integral and minimizing the mean square error (Gelman et al., 2003). Beyond computing Bayesian point-wise estimators, the Bayesian framework also permits to derive precious credibility intervals which can be used to assess the uncertainty associated to the estimation of unknown parameters. These credibility information are particularly important when no ground truth is available about the parameters to infer (e.g., in astrophysics). For instance, Durmus, Moulines, and Pereyra (2018) built upon this framework to assess with high confidence whether a particular structure appearing in a reconstructed tomographic image was indeed present in the original image. Similarly to the MMSE estimator, these intervals are expressed as integrals and write $\int_{\mathcal{C}_\alpha} \pi(\theta|y)\mathrm{d}\theta$ where $\mathcal{C}_\alpha$ is an $(1-\alpha)$ credibility region such that $\mathbb{P}_\pi(\theta \in \mathcal{C}_\alpha) = 1-\alpha$, with $\alpha \in (0,1)$. In most real-world applicative contexts, computing such integrals is not straightforward. One alternative lies in making use of Monte Carlo integration, which consists

in approximating any expectation of the form

$$\mathbb{E}[h(Z)|Z \sim p(z)] = \int h(z)p(z)\mathrm{d}z \tag{0.2}$$

by the empirical average

$$\bar{h} = \frac{1}{N+1}\sum_{n=0}^{N} h(z_n), \tag{0.3}$$

where $\{z_0, \ldots, z_N\}$ is a sample drawn from the distribution $p(z)$ (Robert and Casella, 2004). Obviously, when dealing with Bayesian inference problems, the distribution $p(z)$ of interest is chosen as the targeted posterior distribution $\pi(\theta|y)$.

Markov chain Monte Carlo sampling

This so-called Monte Carlo integration requires the availability of efficient algorithmic schemes able to generate samples from a desired distribution. For simple and univariate probability distributions (e.g., uniform, normal or exponential), generating these samples can be performed via the use of pseudo-random generators combined with transform methods (Devroye, 1986). As a typical example, generating normal random samples can be performed through the Box-Muller transform by exploiting the radial symmetry of the normal distribution (Box and Muller, 1958). For distributions from which it is difficult to simulate via the above approach, some surrogates have been proposed such as (adaptive) rejection sampling (Gilks and Wild, 1992) and (adaptive) importance sampling methods (Bucher, 1988; Geweke, 1989) which sample from a *simpler* instrumental distribution and correct the error with an acceptance or normalization step, respectively. For instance, the standard rejection sampling approach generates samples according to a probability distribution with density p with the following procedure, illustrated in Figure 1:

1. Generate $\theta \sim q$, where q satisfies $p(\theta) \leq Mq(\theta)$ with $M \geq 1$.
2. Let $U \sim \mathcal{U}([0,1])$. Accept θ if $U \leq \frac{p(\theta)}{Mq(\theta)}$.
3. Return to 1. otherwise.

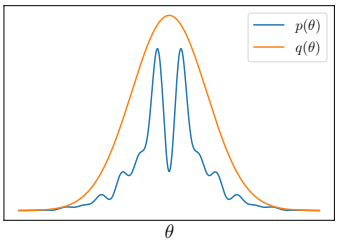


Figure 1: Illustration of the rejection sampling method. Here the instrumental distribution q is a Gaussian from which it is simple to generate samples.

Unfortunately, the aforementioned methods are known to suffer from severe limitations. For instance, it is not always possible to bound the ratio $p(\theta)/q(\theta)$ with a tight constant M for rejection sampling, especially in high dimensional settings (Andrieu et al., 2003).

To cope with these issues, a huge literature dedicated to random variable generation has proposed various Markov chain Monte Carlo (MCMC) algorithms. These algorithms, of which particular instances are the famous Gibbs sampler (Geman and Geman, 1984) and Metropolis-Hastings algorithm (Metropolis and Ulam, 1949; Hastings, 1970), define a wide class of algorithms which allow a Markov chain $\{z_0, \ldots, z_N\}$ to be generated with stationary distribution p (Gilks, Richardson, and Spiegelhalter, 1995). Since Gibbs samplers can be interpreted as special instances of Metropolis ones, we only briefly describe the mechanism of the Metropolis-Hastings

method and refer the interested reader to the textbook by Robert and Casella (2004). Given a complicated target distribution p, this algorithm involves sampling a candidate value according to a *simpler* proposal distribution q before accepting it with some acceptance probability so that the invariant distribution is the target p. A famous instance of Metropolis-Hastings approaches is the random walk algorithm which is based on proposal distributions of the form $\boldsymbol{\theta}_{\text{prop}} = \boldsymbol{\theta}^{(t-1)} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is typically distributed according to a symmetric distribution (e.g., uniform or normal), see Figure 2 for an illustration on a toy example with $\boldsymbol{\epsilon} \sim \mathcal{U}([-\delta, \delta])$. Although this algorithm can be shown to be geometrically ergodic under mild assumptions (Jarner and Hansen, 2000), its *generic* and *myopic* nature often implies a slow convergence and bad mixing properties, especially when the dimension of the problem is large. This drawback can be partially mitigated by adapting the proposal distribution at each iteration (Gilks, Roberts, and Sahu, 1998; Andrieu and Robert, 2001; Haario, Saksman, and Tamminen, 2001; Andrieu and Moulines, 2006).

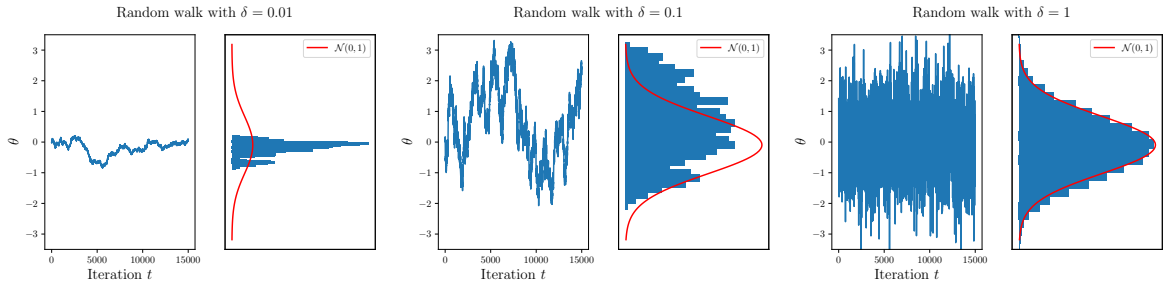


Figure 2: Illustration of the random walk algorithm targeting $\mathcal{N}(0,1)$ with the uniform proposal distribution $\mathcal{U}([-\delta, \delta])$. From left to right, $\delta = 0.01$, $\delta = 0.1$ and $\delta = 1$. When δ is too small (e.g., $\delta = 0.01$), the random-walk algorithm produces highly correlated samples and struggles to explore efficiently the parameter space. On the contrary, setting $\delta = 1$ yields better convergence and mixing properties of the Markov chain.

Guiding Markov chains using first-order information

In order to design appropriate proposal distributions leading to efficient sampling schemes, another line of research has focused on discretizations of continuous-time dynamics which involve a first-order information allowing a better exploration of the state space. These seminal works interestingly lead to hybrid optimization-within-MCMC methods.

Hamiltonian Monte Carlo – For instance, inspired by statistical physics concepts, Hamiltonian Monte Carlo (HMC) algorithms are probably the first technique combining variational optimization (by means of a gradient computation) and Monte Carlo sampling (Duane et al., 1987). The HMC method stands for a specific instance of the Metropolis-Hastings algorithm based on auxiliary variables. The target density $p(\boldsymbol{\theta})$ is augmented by introducing an auxiliary variable $\boldsymbol{\omega}$, called momentum, and such that

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}) = p(\boldsymbol{\theta}) \mathcal{N}(\boldsymbol{\omega} | \mathbf{0}_d, \boldsymbol{\Sigma}). \tag{0.4}$$

The HMC method generates points $(\boldsymbol{\theta}, \boldsymbol{\omega})$ that evolve according to Hamiltonian dynamics given by

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = \nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{\theta}(t), \boldsymbol{\omega}(t)) = \boldsymbol{\Sigma}^{-1} \boldsymbol{\omega}(t) \tag{0.5}$$

$$\frac{\mathrm{d}\boldsymbol{\omega}(t)}{\mathrm{d}t} = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}(t), \boldsymbol{\omega}(t)) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}(t)). \tag{0.6}$$

Since these continuous-time dynamics can be simulated in a few very specific cases, a discretization scheme (e.g., via a leap-frog integrator) is generally used in practice. Because of the discretization scheme, the samples $\{\boldsymbol{\theta}^{(t)}; t \in \mathbb{N}\}$ generated via these discretized Hamiltonian dynamics are distributed according to an approximation of the target $p(\boldsymbol{\theta})$. To obtain samples distributed according to this target, a Metropolis-Hastings correction step is generally used. HMC has proven a remarkable empirical efficiency and as such has been successfully applied in various scenarios (Neal, 2011; Girolami and Calderhead, 2011). Its success is also related to its efficient implementation within popular softwares such as STAN (Carpenter et al., 2017). Interestingly, its theory has been developed and addressed very recently, see for instance the work by Durmus, Moulines, and Eero (2017) and references therein.

Langevin Monte Carlo – Another well-known sampling scheme using gradient information is based on the Langevin diffusion process. It is defined as the solution of the stochastic differential equation

$$\mathrm{d}\boldsymbol{\theta}(t) = \frac{1}{2}\nabla \log p(\boldsymbol{\theta}(t))\mathrm{d}t + \mathrm{d}W(t), \quad \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \tag{0.7}$$

where W stands for the Brownian motion process on $\mathbb{R}^d$. Under some conditions on p (e.g., continuous differentiability), the Langevin diffusion $\boldsymbol{\theta}(t)$ admits an invariant distribution which coincides with p (Roberts and Tweedie, 1996). In practice, a discretization of this diffusion is used (e.g., with the Euler-Maruyama method) and yields the so-called *unadjusted Langevin algorithm* (ULA) defined by the recursion

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \gamma \nabla \log p\left(\boldsymbol{\theta}^{(t-1)}\right) + \sqrt{2\gamma}\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), \tag{0.8}$$

where γ is the step-size associated to the discretization scheme. The bias induced by the latter can be corrected by adding an accept/reject step leading to the Metropolis adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996). These two algorithms recently received a lot of attention, especially since the important work by Durmus and Moulines (2017) who derived non-asymptotic convergence rates for ULA with explicit dependencies to the dimension of the problem, the prescribed precision and regularity constants associated to the target distribution. This work has then been followed by a lot of related contributions (Dalalyan, 2017; Cheng et al., 2018; Chen and Vempala, 2019; Dwivedi et al., 2019).

Exploiting the synergy between Monte Carlo sampling and optimization

Although they have proven to be efficient in lots of applications, standard Hamiltonian and Langevin approaches still suffer from some limitations. They need for instance the continuous differentiability of the target distribution and become computationally prohibitive in big data settings involving a large number of observations (Bardenet, Doucet, and Holmes, 2017). Among others, these two issues have been interestingly addressed by building upon efficient tools and methods used in the optimization literature, drawing new connections between simulation and optimization fields. In the following, we will focus on some of these recent efforts that

have been devoted to cross-fertilize the respective advantages of Monte Carlo and optimization methods.

Proximal Langevin Monte Carlo – Recently, Pereyra (2016) proposed an innovative combination of convex optimization and MCMC algorithms. Capitalizing on the advantages of proximal splitting recently popularized to solve large-scale inference problems (Elad, 2006; Bioucas-Dias and Figueiredo, 2007; Combettes and Pesquet, 2011), the proximal Monte Carlo method permits the sampling from high-dimensional log-concave and potentially non-smooth distributions using the proximity operator (Moreau, 1965). Compared to ULA in (0.8), its Euler discretization scheme writes

$$\boldsymbol{\theta}^{(t)} = \left(1 - \frac{\gamma}{\lambda}\right) \boldsymbol{\theta}^{(t-1)} + \frac{\gamma}{\lambda} \mathrm{prox}_{\lambda \log p}\left(\boldsymbol{\theta}^{(t-1)}\right) + \sqrt{2\gamma}\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), \tag{0.9}$$

where $\mathrm{prox}_{\lambda f}$ is the proximity operator of f. The standard proximal MCMC method requires the availability of the proximal mapping of the log-posterior density. When $\log p = f_1 + f_2$, this mapping is in practice approximated by using a forward-backward splitting scheme. To bypass this issue, Durmus, Moulines, and Pereyra (2018) proposed another proximal MCMC scheme which only needs to have access to the proximal mapping of the non-smooth part of $\log p$. These two works have been generalized in the work by Luu, Fadili, and Chesneau (2020) where the authors in particular relaxed the log-concavity assumption on p.

$$\mathrm{prox}_{\lambda f}(\boldsymbol{\theta}) \coloneqq \underset{\mathbf{z} \in \mathbb{R}^d}{\arg\min} \left\{ f(\mathbf{z}) + \frac{\|\mathbf{z} - \boldsymbol{\theta}\|^2}{2\lambda} \right\}.$$

Stochastic gradient Langevin dynamics – Motivated by large-scale inference, Welling and Teh (2011) considered combining ideas from optimization and simulation in order to fill the gap between these two fields. To that purpose, they proposed to build on *stochastic optimization* (Robbins and Monro, 1951) to scale Langevin Monte Carlo algorithms to tall datasets which involve a very large number of observations (Bardenet, Doucet, and Holmes, 2017). This yields an approach based on so-called *stochastic gradient Langevin dynamics* (SGLD) where the main idea is to approximate the true gradient on the whole dataset of size N with a gradient computed only on a subset of size $n \ll N$. Starting from the ULA recursion (0.8) with $\log p(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p_n(\boldsymbol{\theta})$, the SGLD writes

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \gamma \frac{N}{n} \sum_{i=1}^{n} \nabla \log p_i\left(\boldsymbol{\theta}^{(t-1)}\right) + \sqrt{2\gamma}\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d). \tag{0.10}$$

This implies significant savings in terms of computational time since each iteration now requires $\mathcal{O}(n)$ gradient computations instead of $\mathcal{O}(N)$. For more details about this approach and its extensions, we refer the interested reader to the recent review by Brosse, Moulines, and Durmus (2018).

Majorize-minimize adapted Metropolis–Hastings – Another interesting extension of Langevin Monte Carlo methods based on optimization is the so-called *Majorize-minimize adapted Metropolis–Hastings* introduced in the recent work by Marnissi et al. (2020). The authors proposed to adapt the Gaussian proposal distribution used in ULA and MALA by building on the majorize-minimize (MM) framework (Hunter and Lange, 2004)

to improve the mixing properties of the standard MALA. By adopting an MM quadratic strategy, the Langevin proposal distribution in (0.8) is preconditioned with a positive definite matrix $\mathbf{Q}(\boldsymbol{\theta})$ such that

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \gamma \mathbf{Q}^{-1}\left(\boldsymbol{\theta}^{(t-1)}\right) \nabla \log p\left(\boldsymbol{\theta}^{(t-1)}\right) + \sqrt{2\gamma}\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{0}_d, \mathbf{Q}^{-1}(\boldsymbol{\theta}^{(t-1)})\right). \tag{0.11}$$

By taking into account the local curvature of the posterior density, this adaptive strategy has proven to be successful on various high-dimensional signal and image processing problems.

Gaussian sampling and linear solvers – We end up this short review by pointing out a specific sampling problem which has highly benefited from efficient optimization tools and methods, namely high-dimensional Gaussian sampling. Due to its many nice properties (e.g., infinite divisibility, maximum entropy property or link to the central limit theorem), the Gaussian distribution is ubiquitous. Hence, efficient sampling from the latter is a high-stake problem. In high-dimension, sampling from this distribution raises several important issues which are mainly related to the structure of the covariance (or precision) matrix. To cope with these issues, several methods inspired from the ones used in linear algebra have been developed. For instance, we can cite stochastic adaptations of polynomial and Lanczos approximation approaches (Parker and Fox, 2012; Pereira and Desassis, 2019), conjugate gradient-based samplers (Papandreou and Yuille, 2011; Gilavert, Moussaoui, and Idier, 2015) or Gibbs samplers based on matrix splitting (Fox and Parker, 2017). Some of these approaches will be detailed in Chapter 4.

Remaining challenges to address

This brief overview showed that a lot of approaches have been proposed to improve the efficiency of sampling approaches by cross-fertilizing the mutual benefits of simulation and optimization. Still, some difficult statistical problems remain unsolved. We detail hereafter two of them which will be tackled in this manuscript.

Composite and complicated probability distributions – With the increasing amount and variety of available data and recent advances in specific research fields (e.g., signal and image processing, astrophysics), statistical inference problems become more and more challenging to solve. In particular, this phenomenon arises in many different scenarios when considering the Bayesian framework. The large number of observations involving potential outliers yields complex likelihood functions (e.g., robust losses defined as large sums over the whole training dataset) while the need to encode additional prior information (e.g., non-negativity, spatial, spectral and rank constraints) complicates posterior inference. Indeed, sophisticated and composite prior distributions now often involve non-conjugate, non-separable, non-differentiable and even non-convex terms (Pereyra et al., 2016). These difficulties unfortunately rule out the use of common sampling techniques as the ones reviewed in the previous paragraphs.

Scalable MCMC sampling – Although recent advances in Monte Carlo sampling contributed to decrease the number of iterations required to reach convergence and the associated computational time, MCMC algorithms still remain costly in general. In particular, despite the efforts highlighted in the previous paragraphs, MCMC approaches do not benefit from all the sophisticated tools introduced in optimization which make them very attractive for large-scale inference. Hence, one of the numerous remaining challenges is to carry on contributing to fill the gap between the optimization and stochastic simulation fields in terms of computational cost and scalability.

Contributions and structure of the manuscript

The work presented in this manuscript is an attempt to tackle the aforementioned sampling challenges. Similarly to the works reviewed above, the solutions that are proposed in this manuscript are strongly related to optimization and as such contribute to open new connections between this field and Monte Carlo sampling. The main contributions of this work, divided in chapters, are detailed below.

Chapter 1 proposes a broad and unifying approximate statistical framework, coined *asymptotically exact data augmentation* (AXDA), for inferring unknown quantities in complicated models. Compared to classical data augmentation approaches (van Dyk and Meng, 2001), AXDA circumvents the art of finding the exact augmented model associated to each specific situation to build efficient inference algorithms. AXDA considers an approximate model whose bias is assessed under various assumptions. Interestingly, it can be related to *approximate Bayesian computation* approaches (Marin et al., 2012) and can benefit from its existing and efficient algorithms.

Chapter 2 presents a specific MCMC algorithm called *split Gibbs sampler* (SGS) dedicated to sample from an AXDA model. We show that SGS shares strong connections with popular optimization approaches such as quadratic penalty methods and the alternating direction method of multipliers (ADMM). Similarly to these deterministic methods, SGS benefits from interesting properties. It stands for a divide-to-conquer approach, is simple, scalable in distributed environments, and its empirical performances compete with (and sometimes even improve upon) state-of-the-art approaches.

Chapter 3 proposes a detailed theoretical study of the convergence properties of SGS. Under regularity conditions, we establish explicit and non-asymptotic convergence rates for this scheme using Ricci curvature and coupling ideas. Combined with bias bounds on the AXDA approximation, we provide complexity results for SGS with explicit dependencies with respect to the dimension of the problem, the prescribed precision and regularity constants associated to the target posterior distribution. The work presented in this chapter is the result of an international collaboration with researchers from the University of Oxford (Arnaud Doucet) and

the University of Edinburgh (Daniel Paulin). This collaboration started in 2019 during a 2-month visiting period in Arnaud Doucet's research group.

Chapter 4 is dedicated to high-dimensional Gaussian sampling, a step that often arises when considering SGS. In this chapter, we first review the main Gaussian sampling techniques based on MCMC samplers. We would like to emphasize that such a review encompassing recent Gaussian sampling approaches proposed by distinct communities does not exist. On top of that review, we propose to shed new light on most of these techniques by embedding them into a unifying framework based on a stochastic counterpart of the celebrated proximal point algorithm.

Chapter 5 concludes the study of the proposed AXDA framework by analyzing it from an optimization point of view. To that purpose, this chapter focuses on the negative log density (also called potential function) coming from AXDA densities. We show that this potential function defines a class of smooth envelope functions which converge to the famous Moreau envelope in a limiting case. Combined with the results in Chapter 1, this chapter allows to have a complete understanding of the approximation involved in AXDA.

All the proofs associated to the results shown in this manuscript are postponed to Appendices associated to each chapter.

For sake of reproducible research, the code associated to the numerical results presented in our research works is available online at

https://github.com/mvono

List of publications

Most of the work presented in this manuscript has been published or is currently under review for publication.

Submitted

- M. Vono, D. Paulin, and A. Doucet (2019). "Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting." In 2nd round of review, *Journal of Machine Learning Research.* arXiv: 1905.11937
- M. Vono, N. Dobigeon, and P. Chainais (2020b). "High-dimensional Gaussian sampling: A review and a unifying approach based on a stochastic proximal point algorithm." In 1st round of review, *SIAM Review.* arXiv: 2010.01510

International journals

- M. Vono, N. Dobigeon, and P. Chainais (2020a). "Asymptotically exact data augmentation: models, properties and algorithms." *Journal of Computational and Graphical Statistics (in press).* doi:10.1080/10618600.2020.1826954

M. Vono, N. Dobigeon, and P. Chainais (2019a). "Split-and-augmented Gibbs sampler - Application to large-scale inference problems." *IEEE Transactions on Signal Processing* 67 (6): 1648–1661. doi:10.1109/TSP.2019.2894825

International conferences

M. Vono, N. Dobigeon, and P. Chainais (2018). "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler." In *IEEE International Workshop on Machine Learning for Signal Processing*. doi:10.1109/MLSP.2018.8516963

M. Vono, N. Dobigeon, and P. Chainais (2019c). "Efficient sampling through variable splitting-inspired Bayesian hierarchical models." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/ICASSP.2019.8682982

M. Vono, N. Dobigeon, and P. Chainais (2019b). "Bayesian image restoration under Poisson noise and log-concave prior." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/ICASSP.2019.8683031

National conferences

M. Vono, N. Dobigeon, and P. Chainais (2019d). "Modèles augmentés asymptotiquement exacts." In *Proc. of GRETSI*

M. Vono, N. Dobigeon, and P. Chainais (2019e). "Un modèle augmenté asymptotiquement exact pour la restauration bayésienne d'images dégradées par un bruit de Poisson." In *Proc. of GRETSI*

List of publications not related to this manuscript

During this Ph.D. work, I was also involved in the ORION-B consortium which gathers astrophysicists, statisticians and engineers to better understand the formation of stars in molecular clouds. More information are available on the website of this project: https://www.iram.fr/~pety/ORION-B/. As part of this project, I participed to annual meetings and the following work have been published.

M. Vono et al. (2019). "A fully Bayesian approach for inferring physical properties with credibility intervals from noisy astronomical data." In *Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing*, 1–5. doi:10.1109/WHISPERS.2019.8920859

A. Roueff et al. (2020). "C18O, 13CO, and 12CO abundances and excitation temperatures in the Orion B molecular cloud: An analysis of the precision achievable when modeling spectral line within the Local Thermodynamic Equilibrium approximation." *Astronomy & Astrophysics (in press)*. arXiv: 2005.08317

A journal paper extending the results of our conference paper is planned to be submitted before the end of the year.

Academic visits and invited talks

Apart from publications in conferences and journals, I had the honor to spend few months abroad and to give invited talks. These experiences which contributed to disseminate and to improve the work presented in this manuscript are detailed below.

- February 2019 - April 2019. **Research visiting scholar** at the Department of Statistics of the **University of Oxford**, invited by Arnaud Doucet (Professor of Statistics). This visit leads to a research work submitted to an international journal (Vono, Paulin, and Doucet, 2019).
- December 2019. **Invited talk** given at the **University of Hong Kong** (Shenzhen, China) as part of the Workshop on optimization, probability and simulation (WOPS) organized by the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS).
- February 2020. **Invited talk** given at the **Heriot-Watt University** (Edinburgh, UK) as part of the Actuarial Mathematics & Statistics seminars organized by the School of Mathematics.
- Autumn 2020. **Invited talk** as part of the Bayesian Machine Learning at Scale seminar series jointly organized by **Criteo AI Lab** and Nicolas Chopin (ENSAE). Website of the event: https://ailab.criteo.com/laplaces-demon-bayesian-machine-learning-at-scale.

1 Asymptotically exact data augmentation

"It's often when you're talking over things that you seem to see your way clear. Your mind gets made up for you sometimes without your knowing how it's happened. Talking leads to a lot of things one way or another."

— Agatha Christie, *The A.B.C. Murders*

Chapter contents

The purpose of this chapter is to lay the foundations of a broad, rigorous and systematic framework, coined *asymptotically exact data augmentation* (AXDA), for inferring unknown parameters in complex statistical models. As its name implies, this framework constitutes a class of approximate augmented models. In both this chapter and Chapter 2, we show that such an approximation provides an answer to the main issue of *data augmentation* approaches, namely the art of finding the exact augmented model which yields efficient inference algorithms (van Dyk and Meng, 2001). Remarkably, AXDA models may also inherit interesting properties such as sophisticated computational approaches from the *approximate Bayesian computation* (ABC) literature (Sisson, Fan, and Beaumont, 2018b), scalability in distributed architectures and robustness. Overall, at the price of an approximation which comes with theoretical guarantees, AXDA approaches will appear to be a general and efficient way to conduct simple inference in a wide variety of large-scale problems.

Depending on the context, the name *data augmentation* may have different meanings. In the machine and deep learning community, this term refers to the process of increasing the size of the training data by manipulating the original data (Goodfellow, Bengio, and Courville, 2016). Figure 1.1 illustrates this idea with an image from the CIFAR-10 dataset (Krizhevsky, 2009). In this manuscript, this term instead covers the whole range of methods for constructing iterative optimization or simulation-based algorithms via the introduction of auxiliary (also called latent) variables (van Dyk and Meng, 2001).

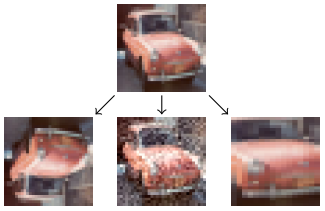


Figure 1.1: An illustration of data augmentation in machine learning. The top image, taken from the CIFAR-10 dataset, represents an old car. The three images at the bottom stand for new images built from the original one via simple transformations: rotation, noise addition and zoom-in.

The AXDA framework is introduced in Section 1.1. Then, Section 1.2 revisits some already-proposed special instances of AXDA models to exhibit interesting properties which can be generally inherited by the proposed framework. In Section 1.3, we assess quantitatively the bias induced by resorting to this class of approximate models. These results are finally illustrated numerically in Section 1.4. Proofs are collected in Appendices at the end of the manuscript.

The major part of the material of this chapter is currently in second revision in an international journal and has been presented at a national conference. Some of the theoretical results shown in Section 1.3 are part of a complementary work submitted to another international journal:

M. Vono, N. Dobigeon, and P. Chainais (2020a). "Asymptotically exact data augmentation: models, properties and algorithms." *Journal of Computational and Graphical Statistics (in press)*. doi:10.1080/10618600.2020.1826954

M. Vono, D. Paulin, and A. Doucet (2019). "Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting." In 2nd round of review, *Journal of Machine Learning Research*. arXiv: 1905.11937

M. Vono, N. Dobigeon, and P. Chainais (2019d). "Modèles augmentés asymptotiquement exacts." In *Proc. of GRETSI*

1.1 The approximate distribution

This section introduces the AXDA framework and details how to build its approximate augmented models in a systematic and general way. For sake of simplicity, with little abuse, we shall use the same notations for a probability distribution and its associated probability density function (pdf). In the sequel, we assume that all the probability distributions admit a pdf with respect to (w.r.t.) the Lebesgue measure.

1.1.1 *Motivations*

We are interested in performing the statistical inference of an unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ by (partly or fully) relying on a probability distribution having a proper density π which writes

$$\pi(\boldsymbol{\theta}) = \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^d} e^{-f(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}}, \quad \text{or} \quad \pi(\mathbf{y}|\boldsymbol{\theta}) = \frac{e^{-f(\mathbf{y};\boldsymbol{\theta})}}{\int_{\mathbb{R}^n} e^{-f(\mathbf{y};\boldsymbol{\theta})} \mathrm{d}\mathbf{y}}, \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ refers to a set of observations. In both cases, the potential f taking values in the extended real line $\mathbb{R} \cup \{+\infty\}$ is a proper and lower semi-continuous function. This extended image domain will allow us to consider, for instance, indicator functions of convex sets (Boyd and Vandenberghe, 2004, Section 3.1.2) which are ubiquitous in statistical signal processing and machine learning. For the sake of generality, notice that π in (1.1) 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 density $\pi(\boldsymbol{\theta})$ or its posterior density $\pi(\boldsymbol{\theta}) \coloneqq \pi(\boldsymbol{\theta}|\mathbf{y})$. Depending on the problem, we also allow π to stand for a likelihood function $\pi(\mathbf{y}|\boldsymbol{\theta})$. We will work under this slightly abusive convention and write explicitly the form of π 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 using π in (1.1) is difficult because intractable or computationally prohibitive. The first difficulty arises for instance in maximum likelihood estimation problems (Filstroff, Lumbreras, and Févotte, 2018) while the second one is generally related to slow-mixing Markov chains when Monte Carlo sampling methods are

In the sequel, the adverb *partly* will refer to the case where the inference is performed within a Bayesian framework and π stands for either the likelihood or the prior. On the other hand, the inference will *fully* rely on π when the latter is a posterior density in a Bayesian setting or a likelihood in a frequentist one.

Since the beginning of this section, note that the adjective *proper* has been used twice and with two different meanings. In probability theory, a positive-valued pdf π is proper if $\int \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$. In analysis, a function f is said to be proper if there exists $\mathbf{x}_0$ such that $f(\mathbf{x}_0) < \infty$ and $f(\mathbf{x}) > -\infty$ for every $\mathbf{x} \in \mathrm{dom} f$, where $\mathrm{dom} f = \{\mathbf{x} \mid f(\mathbf{x}) < \infty\}$.

The notation $\coloneqq$ means "equal to, by definition".

employed (Duane et al., 1987; Edwards and Sokal, 1988). To overcome these issues, an option is to rely on exact data augmentation (DA) which introduces some auxiliary variables stacked into a vector $\mathbf{z} \in \mathbb{R}^k$ (Tanner and Wong, 1987). Then, it defines a joint pdf $\pi(\boldsymbol{\theta}, \mathbf{z})$ that is simpler to handle and such that the marginal density of $\boldsymbol{\theta}$ under this joint model coincides with the original one, i.e.,

$$\int_{\mathbb{R}^k} \pi(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} = \pi(\boldsymbol{\theta}). \tag{1.2}$$

Throughout this manuscript, whenever appropriate, all equalities or inequalities such as (1.2) are understood to hold almost surely w.r.t. an appropriate dominating measure.

The popularity of DA approaches can be traced back to the use of latent variables in likelihood inference in the 1960s, and to the seminal paper by Dempster, Laird, and Rubin (1977) on the expectation-maximization algorithm. Since then, 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 (Celeux et al., 2001; Doucet, Godsill, and Robert, 2002; Marnissi et al., 2018).

Nevertheless, these approaches have several limitations. Indeed, finding a convenient form for the augmented density in order to satisfy (1.2) while leading to efficient algorithms generally requires some expertise and can even be impossible in some cases (Geman and Yang, 1995; van Dyk and Meng, 2001). For instance, the mixture representation of a binomial likelihood function based on the Pólya-Gamma distribution has been used to derive a promising Gibbs sampler for logistic regression problems (Polson, Scott, and Windle, 2013). Nevertheless, although this algorithm has been shown to be uniformly ergodic w.r.t. the TV distance, the best known explicit result for its ergodicity constant depends exponentially on the number of observations n and on the dimension of the regression coefficients vector d, see the work by Choi and Hobert (2013).

To tackle these limitations, some techniques have been proposed such as partial decoupling (Higdon, 1998) or the use of a working parameter (Meng and van Dyk, 1997, 1998; Liu and Wu, 1999). In the sequel, we propose to take a new route by relaxing the constraint (1.2) and considering an *approximate* augmented 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. This so-called AXDA framework, which embeds approximate DA models controlled by a positive scalar parameter ρ, is presented in Section 1.1.2. These models become asymptotically exact when ρ 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 quantitatively in Section 1.3 with non-asymptotic theoretical results.

1.1.2 *Model*

Instead of searching for an exact data augmentation scheme (1.2), some auxiliary variable $\mathbf{z} \in \mathbb{R}^k$ 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 scalar tolerance

parameter $\rho > 0$ and such that the associated marginal density defined by

$$\pi_\rho(\boldsymbol{\theta}) = \int_{\mathbb{R}^k} \pi_\rho(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z}, \tag{1.3}$$

satisfies the following property:

Property 1. *For all* $\boldsymbol{\theta} \in \mathbb{R}^d$, $\lim_{\rho \to 0} \pi_\rho(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

The prerequisite of Property 1 is essential since, similarly to the condition (1.2) for DA, the auxiliary variable $\mathbf{z}$ is introduced here for computational purposes and should not alter significantly the initial density π. By applying Scheffé's lemma (Scheffé, 1947), this property yields the convergence in total variation as detailed in the following corollary:

Corollary 1. *Under Property 1,* $\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \to 0$ *as* $\rho \to 0$.

The quantity $\|\mu - \nu\|_{\mathrm{TV}}$ stands for the total variation distance between the measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and is defined by

$$\sup_{f \in \mathbb{M}(\mathbb{R}^d), \|f\|_\infty \leq 1} \left| \int_{\mathbb{R}^d} f(\boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{\theta}) - \int_{\mathbb{R}^d} f(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta}) \right|,$$

where $\mathbb{M}(\mathbb{R}^d)$ denotes the set of all Borel measurable functions f on $\mathbb{R}^d$.

A natural question is: how to choose the augmented density in (1.3) such that Property 1 is met? In this paper, we assume that $\mathbf{z}$ and $\boldsymbol{\theta}$ live in the same space, that is $k = d$, and investigate AXDA schemes associated to an initial density (1.1) and defined by the approximate augmented density

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}) = \frac{1}{Z_\rho} \pi(\mathbf{z}) \kappa_\rho(\mathbf{z}; \boldsymbol{\theta}), \tag{1.4}$$

where $Z_\rho = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \pi(\mathbf{z}) \kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) \mathrm{d}\mathbf{z} \mathrm{d}\boldsymbol{\theta} > 0$ stands for a normalizing constant and κ_ρ is such that (1.4) defines a proper joint density. Since π is a proper density, a simple sufficient assumption to ensure this property is that $\int_{\mathbb{R}^d} \kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty$.

Remark 1. *When* π *stands for a product of* $b \geq 1$ *densities, that is* $\pi = \prod_{i=1}^b \pi_i$, *the proposed approximate model can be naturally generalized to* $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b}) \propto \prod_{i=1}^b \pi_i(\mathbf{z}_i) \kappa_\rho(\mathbf{z}_i; \boldsymbol{\theta})$. *Such a generalization will for instance be considered in Sections 1.2.1 and 1.2.2.*

A sufficient condition to satisfy Property 1 is to require that the sequence $\kappa_\rho(\cdot; \boldsymbol{\theta})$ weakly converges towards the Dirac distribution concentrated at the point $\boldsymbol{\theta}$, denoted by $\delta(\cdot - \boldsymbol{\theta})$, as $\rho \to 0$, that is for any bounded and continuous function ψ we have

$$\int_{\mathbb{R}^d} \psi(\mathbf{z}) \kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) \mathrm{d}\mathbf{z} \xrightarrow[\rho \to 0]{} \int_{\mathbb{R}^d} \psi(\mathbf{z}) \delta(\mathbf{z} - \boldsymbol{\theta}) \mathrm{d}\mathbf{z} = \psi(\boldsymbol{\theta}). \tag{1.5}$$

This convergence is illustrated in the scalar case on Figure 1.2 for $\kappa_\rho(z; \theta) = \mathcal{N}(z; \theta, \rho^2)$.

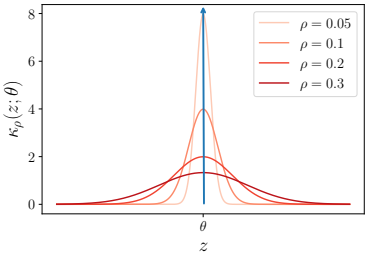


Figure 1.2: Example of a sequence $\kappa_\rho(z; \theta)$ weakly converging towards $\delta(z - \theta)$ as $\rho \to 0$.

> In the sequel, we will call AXDA the family of approaches based on the augmented model defined by (1.4) and satisfying the weak convergence property (1.5).

One of the aims of introducing the proposed joint model (1.4) is to avoid a case-by-case search of an appropriate augmented approach. Hence, although there might exist other marginal densities π_ρ satisfying Property 1, we restrict our analysis to the so-called AXDA framework. The

following paragraphs detail two possible ways to build models belonging to this framework.

AXDA using standard kernels – One possibility to construct the sequence κ_ρ is to consider a kernel K, that is a positive function such that $\int_{\mathbb{R}^d} K(\mathbf{u})\mathrm{d}\mathbf{u} = 1$ and $K(-\mathbf{u}) = K(\mathbf{u})$, for all $\mathbf{u} \in \mathbb{R}^d$. Based on the latter we define (Dang and Ehrhardt, 2012), for all $\mathbf{z}, \boldsymbol{\theta} \in \mathbb{R}^d$,

$$\kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) = \rho^{-d} K\left(\frac{\mathbf{z} - \boldsymbol{\theta}}{\rho}\right). \tag{1.6}$$

Table 1.1 lists some classical examples of kernels K which are not necessarily compactly supported. For the sake of simplicity, we only define univariate versions of them but they can obviously be generalized in higher dimensions. Figure 1.3 illustrates these kernels. Standard kernels have already been used in the statistical community to define approximate probability density functions of the form (1.4). For instance, *noisy* ABC methods (Fearnhead and Prangle, 2012; Wilkinson, 2013) build upon the same type of approximation detailed in this paragraph. As such, they can be related to the proposed AXDA framework. The benefits of this relation will be detailed in Section 1.2.4.

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

Table 1.1: Examples of classical kernels K that can be used to define κ_ρ.

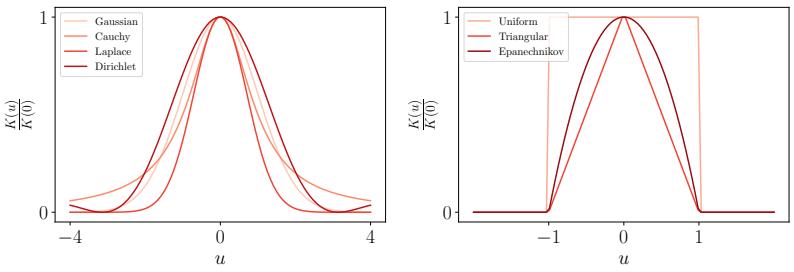


Figure 1.3: (left) Normalized non compactly-supported kernels; (right) normalized compactly-supported kernels detailed in Table 1.1.

AXDA using divergences – Beyond the kernels listed in Table 1.1 but motivated by the same idea of measuring the discrepancy between the latent variable $\mathbf{z}$ and the initial one $\boldsymbol{\theta}$, another general strategy to derive κ_ρ is to build on divergence functions $\phi(\mathbf{z}, \boldsymbol{\theta})$ widely used in the optimization literature (Ben-Tal, Margalit, and Nemirovski, 2001; Beck and Teboulle, 2003; Duchi et al., 2012; Krichene, Bayen, and Bartlett, 2015). For all $\mathbf{z}, \boldsymbol{\theta} \in \mathbb{R}^d$, if we define

$$\kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) \propto_{\mathbf{z}} \exp\left(-\frac{\phi(\mathbf{z}, \boldsymbol{\theta})}{\rho}\right), \tag{1.7}$$

where ϕ is a strictly convex function w.r.t. $\mathbf{z}$ admitting a unique minimizer $\mathbf{z}^* = \boldsymbol{\theta}$, then under mild differentiability assumptions on ϕ, one can show that κ_ρ satistifies the weak convergence property (1.5) underlying an AXDA model (Fellows et al., 2019, Theorem 1). An archetypal example is the scenario where $\phi \coloneqq d_\psi : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty]$ stands for the Bregman divergence (Bregman, 1967) w.r.t. some continuously differentiable and strictly convex function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ on $\mathrm{int}(\mathrm{dom}\psi)$, defined by

A function $f : \mathbb{R}^d \to \mathbb{R}$ is strictly convex if for all $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \in \mathbb{R}^d$ and $t \in (0,1)$, we have $f(t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2) < tf(\boldsymbol{\theta}_1) + (1-t)f(\boldsymbol{\theta}_2)$.

$$d_\psi(\mathbf{z}, \boldsymbol{\theta}) = \begin{cases} \psi(\mathbf{z}) - \psi(\boldsymbol{\theta}) - \langle \nabla\psi(\boldsymbol{\theta}), \mathbf{z} - \boldsymbol{\theta} \rangle & \forall (\mathbf{z}, \boldsymbol{\theta}) \in \mathrm{dom}\psi \times \mathrm{int}(\mathrm{dom}\psi), \\ \infty & \text{otherwise,} \end{cases} \tag{1.8}$$

where $\mathrm{int}(\mathrm{dom}\psi)$ denotes the interior of the domain of ψ. Univariate examples of such divergence functions with their respective domains are listed in Table 1.2. The use of divergence functions to define probability distributions is not new and does not come as a surprise. Indeed, there exists a strong relationship between Bregman divergences and the exponential family, which gathers many of the most commonly-used distributions (Azoury and Warmuth, 2001; Banerjee et al., 2005; Févotte, Bertin, and Durrieu, 2009).

loss function ψ	$\psi(\theta)$	$\mathrm{dom}\psi$	$d_\psi(z, \theta)$
Energy	θ^2	$\mathbb{R}$	$(z-\theta)^2$
Fermi-Dirac entropy	$\theta\log(\theta) + (1-\theta)\log(1-\theta)$	$[0,1]$	$z\log\left(\frac{z}{\theta}\right) + (1-z)\log\left(\frac{1-z}{1-\theta}\right)$
Boltzmann–Shannon	$\theta\log(\theta)$	$[0,1]$	$z\log\left(\frac{z}{\theta}\right) - z + \theta$

Table 1.2: Examples of potentials $\phi \coloneqq d_\psi$ that can be used to define an appropriate density κ_ρ verifying Property 1.

1.2 Benefits of AXDA by revisiting existing models

Before assessing the bias of AXDA models with quantitative results, this section proposes to review some important state-of-the-art works from the AXDA perspective described in Section 1.1. We do not pretend to give new insights about these approaches. We rather use them to exhibit 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.

1.2.1 *Tractable posterior inference*

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

To this purpose, we consider the case where the posterior distribution is intractable. Such an issue for instance appears when this posterior involves a constraint on some set (Liechty, Liechty, and Müller, 2009), a non-standard potential function such as the total variation semi-norm (Vono, Dobigeon, and Chainais, 2019a) 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

to a hierarchical Bayesian model. In those cases, the AXDA framework 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 for instance Chapter 2. The following example, derived from Holmes and Mallick (2003), illustrates this idea.

Example.

Let $\mathbf{y} \in \mathbb{R}^n$ be a vector of observations and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ a design matrix filled with covariates. We consider a generalized non-linear model which writes

$$y_i|\boldsymbol{\theta} \sim \pi(y_i \mid g^{-1}(h(\mathbf{x}_i, \boldsymbol{\theta})), \sigma^2), \quad \forall i \in [n], \tag{1.9}$$
$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{0}_d, \nu^2 \mathbf{I}_d), \tag{1.10}$$

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

$$h(\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^{d} \theta_j B(\mathbf{x}_i, \mathbf{k}_j), \tag{1.11}$$

where $B(\cdot, \mathbf{k}_j)$ 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. In order to mitigate this issue, Holmes and Mallick (2003) proposed to rely on an additional level which boils down to considering the approximate model (1.4) applied to each individual contribution to the likelihood $\pi(y_i|\boldsymbol{\theta})$, for $i \in [n]$. More specifically, the aforementioned authors treated the non-linear predictor h as a Gaussian random latent variable which leads to the approximate model

$$y_i|z_i \sim \pi(y_i \mid g^{-1}(z_i), \sigma^2), \quad \forall i \in [n], \tag{1.12}$$
$$z_i|\boldsymbol{\theta} \sim \mathcal{N}(z_i \mid h(\mathbf{x}_i, \boldsymbol{\theta}), \rho^2), \quad \forall i \in [n], \tag{1.13}$$
$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{0}_d, \nu^2 \mathbf{I}_d). \tag{1.14}$$

Here, AXDA has been applied only to the likelihood function with κ_ρ chosen as the univariate normal distribution (1.13) leading to a smoothed likelihood function. Actually, note that a slight generalization of AXDA has been applied by "replacing" $h(\mathbf{x}_i, \boldsymbol{\theta})$, instead of just $\boldsymbol{\theta}$, by the latent variable z_i. The main advantage of relying on such a model is that the posterior conditional distribution of $\boldsymbol{\theta}$ 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, Scott, and Windle, 2013).

Beyond the widely-used Gaussian choice for κ_ρ (Holmes and Mallick, 2003; Liechty, Liechty, and Müller, 2009; Barbos et al., 2017; Vono, Dobigeon, and Chainais, 2019a), more general AXDA approaches can be built by taking inspiration from these works. To this purpose, we recommend to adaptively set κ_ρ 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 $\phi \coloneqq d_\psi$ (see Section 1.1.2) 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).

1.2.2 *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's 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

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \propto \nu(\boldsymbol{\theta}) \prod_{b=1}^{B} \prod_{i \in \text{node } b} \exp\left(-f_i(y_i; h(\mathbf{x}_i, \boldsymbol{\theta}))\right). \tag{1.15}$$

Such models classically appear in statistical machine learning when generalized linear models (GLMs) are considered (Dobson and Barnett, 2008). In these cases, $h(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{x}_i^T\boldsymbol{\theta}$ and an archetypal example is the logistic regression problem. It assumes that the observed binary variables $\mathbf{y} \in \{0,1\}^n$ follow the Bernoulli distribution $\mathcal{B}(\sigma(\mathbf{x}_i^T\boldsymbol{\theta}))$ where $\sigma(\cdot)$ is the logistic link. This leads to the posterior distribution with density

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \propto \nu(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left(-f_i(y_i; \mathbf{x}_i^T\boldsymbol{\theta})\right), \tag{1.16}$$

where $f_i(y_i; \mathbf{x}_i^T\boldsymbol{\theta}) = -y_i \log(\sigma(\mathbf{x}_i^T\boldsymbol{\theta})) - (1-y_i)\log(1-\sigma(\mathbf{x}_i^T\boldsymbol{\theta}))$. The posterior (1.16) can be re-written as in (1.15) by simply gathering the indices $i \in [n]$ associated to data which belong to the same node b.

Due to the distributed environment, sampling efficiently from (1.15) 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 (1.15). Nonetheless, the averaging schemes used to combine the local chains might lead to poor approximations when π is high-dimensional and non-Gaussian, see the work by Rendell et al. (2018) for a comprehensive review. Instead, considering a special instance of AXDA circumvents the previously mentioned drawbacks. It consists in introducing local auxiliary variables on each node such that

Similarly to the example of Section 1.2.1, note that AXDA has been applied to the scalar product $\mathbf{x}_i^T\boldsymbol{\theta}$ instead of $\boldsymbol{\theta}$.

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, \mathbf{X}) \propto \nu(\boldsymbol{\theta}) \prod_{b=1}^{B} \prod_{i \in \text{node } b} \exp\left(-f_i(y_i; z_i)\right) \kappa_\rho(z_i; \mathbf{x}_i^T \boldsymbol{\theta}). \quad (1.17)$$

Let choose κ_ρ to be log-concave w.r.t. z_i and such that $\nu(\boldsymbol{\theta})$ stands for a conjugate prior for κ_ρ. The posterior distribution of the auxiliary variables z_i conditionally to $\boldsymbol{\theta}$ only depends on the data y_i available at a given node. Based on this nice property, the joint posterior can be sampled efficiently thanks to the separability of the posterior distribution of the auxiliary variables conditioned upon $\boldsymbol{\theta}$. Indeed, the conditional distribution of z_i being univariate and log-concave, sampling from it can be done efficiently and in a distributed manner with (adaptive) rejection sampling (Gilks and Wild, 1992). On the other hand, the choice of κ_ρ leads to a standard conditional posterior for $\boldsymbol{\theta}$ which can be sampled with off-the-shelf techniques. 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).

1.2.3 *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 cope with model misspecification by describing 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 by Wang and Blei (2018). This "localization" idea is illustrated in Figure 1.4. Many of these models 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

$$\pi(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|\boldsymbol{\theta}), \quad (1.18)$$

where $\boldsymbol{\theta}$ is a common parameter. Applying AXDA as described in Section 1.1 by introducing n auxiliary variables stacked into the vector $\mathbf{z}_{1:n}$ leads to the augmented likelihood

$$\pi_\rho(\mathbf{y}, \mathbf{z}_{1:n}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \pi(y_i|\mathbf{z}_i) \kappa_\rho(\mathbf{z}_i; \boldsymbol{\theta}). \quad (1.19)$$

The statistical model defined by (1.19) implies a hierarchical Bayesian model similar to the localized one depicted on Figure 1.4 and corresponds in general to an approximation of the initial one, see the following example.

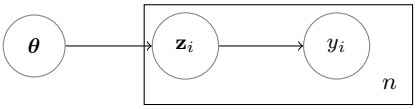


Figure 1.4: Concept of localization: for $i \in [n]$, each observation y_i is assumed to have been generated from a local model depending on a latent variable $\mathbf{z}_i$.

Example.

Assume that for all $i \in [n]$, $\pi(y_i|\boldsymbol{\theta}) = \mathcal{B}\left(\sigma\left(\mathbf{x}_i^T\boldsymbol{\theta}\right)\right)$, where $\mathcal{B}$ stands for the Bernoulli distribution, 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 $\mathcal{B}\left(\sigma\left(\mathbf{x}_i^T\mathbf{z}_i\right)\right)$ associated to an auxiliary parameter $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\theta}, \rho^2\mathbf{I}_d)$. In this case, $\kappa_\rho(\mathbf{z};\boldsymbol{\theta}) = \prod_{i=1}^n \mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\theta}, \rho^2\mathbf{I}_d)$. This model has for example been considered by Rendell et al. (2018) while another approximate logistic regression model has been derived by Vono, Dobigeon, and Chainais (2018).

Beyond the convenient Gaussian prior κ_ρ advocated by Wang and Blei (2018), the choice of κ_ρ through its potential ϕ, see Section 1.1.2, can be motivated by robust loss functions used in the statistical machine learning literature such as the absolute or Huber losses (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 κ_ρ to be the distribution of the modeling error which could be adjusted thanks to some expertise.

1.2.4 *Inheriting sophisticated inference schemes from ABC*

Finally, this section completes the observation made in Section 1.1.2 and shows that AXDA approaches, by sharing strong connections with ABC ones, might inherit sophisticated algorithms to sample from the posterior distribution derived from (1.4).

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, Fan, and Beaumont (2018b) 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 π by considering the pseudo-likelihood with density

$$\pi_\rho(\mathbf{y}|\boldsymbol{\theta}) \coloneqq \int_{\mathbb{R}^n} \pi_\rho(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\mathrm{d}\mathbf{z} = Z_\rho^{-1} \int_{\mathbb{R}^n} \pi(\mathbf{z}|\boldsymbol{\theta})\kappa_\rho(\mathbf{z};\mathbf{y})\mathrm{d}\mathbf{z}. \tag{1.20}$$

When n is large, the high-dimensional observation vector $\mathbf{y}$ is replaced by a low-dimensional summary statistic $\mathbf{s}(\mathbf{y})$. In this case, the latent variable becomes $\mathbf{s}(\mathbf{z})$.

This density has exactly the same formulation as the one defined in (1.4) 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. (2018), one can derive efficient algorithms for AXDA from the ABC framework. For instance, Rendell et al. (2018) recently built on the works by Beaumont, Zhang, and Balding (2002) and Del Moral, Doucet, and Jasra (2012) in the ABC context to propose a bias correction approach and a sequential Monte Carlo algorithm avoiding the tuning of the tolerance parameter ρ. Obviously, many other inspirations from ABC can be considered, such as the parallel tempering approach proposed by Baragatti, Grimaud, and Pommeret (2013) among others, to make the inference from an AXDA model more flexible and efficient.

1.3 Theoretical guarantees

By building on existing approaches, Section 1.2 showed that the AXDA framework can be used in quite general and different settings depending on ones motivations. In order to further promote the use of such an approximate framework, 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 and credibility regions are derived. The proofs are gathered in Appendices at the end of the manuscript.

1.3.1 *Results for standard kernels*

In this section, we consider the case

$$\kappa_\rho(\mathbf{z};\boldsymbol{\theta}) = \rho^{-d} K\left(\frac{\boldsymbol{\theta}-\mathbf{z}}{\rho}\right), \tag{1.21}$$

where K is a kernel, see (1.6). Under this model and based on convolution properties, the following results hold.

Proposition 1. *Let $\pi \in L^1$. The marginal with density π_ρ in (1.3) has the following properties.*

We say that a function $f \in L^1$ if

$$\int |f| \mathrm{d}\mu < \infty,$$

where μ is the Lebesgue measure.

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

$$\mathbb{E}_{\pi_\rho}(\theta) = \mathbb{E}_{\pi}(\theta)$$
$$\mathrm{var}_{\pi_\rho}(\theta) = \mathrm{var}_{\pi}(\theta) + \mathrm{var}_{\kappa_\rho}(\theta).$$

The notation $\mathrm{supp(h)} = \{\mathbf{x} \in \mathcal{X} \mid h(\mathbf{x}) \neq 0\}$ refers to the support of a function $h : \mathcal{X} \to \mathbb{R}$.

(ii) $\mathrm{supp}(\pi_\rho) \subseteq C$ where C is the closure of $\{\mathbf{x}+\mathbf{z}; \mathbf{x} \in \mathrm{supp}(\pi), \mathbf{z} \in \mathrm{supp}(\kappa_\rho)\}$.

(iii) If both π and κ_ρ are log-concave, then π_ρ 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 π_ρ is infinitely differentiable w.r.t. $\boldsymbol{\theta}$.

Proof. See Appendix A.1. □

Proposition 1 permits to draw several conclusions about the inference based on π_ρ. Firstly, the infinite differentiability of π_ρ shown in Property *(iv)* implies that it stands for a smooth approximation of π and might ease the inference. Secondly, Property *(i)* of Proposition 1 is reassuring regarding the inference task. Indeed, if π stands for a prior distribution, then considering the approximation π_ρ simply corresponds to a more diffuse prior knowledge around the same expected value, see Section 1.4.2. Thus, more weight will be given to the likelihood if a posterior distribution is derived with this prior. On the other hand, if π stands for a likelihood, then considering the approximation π_ρ 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 1.2.3.

We now provide quantitative bounds on the approximation implied by considering the approximate marginal density π_ρ instead of π. Under mild assumptions on the kernel K, Proposition 2 gives a simple and practical upper bound on the p-Wasserstein distance between π_ρ and π.

For $p \geq 1$, the p-Wasserstein distance between π and π_ρ, raised to the power p, is defined by

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

where $\Gamma(\pi_\rho, \pi)$ is the set of probability distributions $\mu(\boldsymbol{\theta}, \mathbf{z})$ with marginals π_ρ and π w.r.t. $\boldsymbol{\theta}$ and $\mathbf{z}$, respectively.

Proposition 2. *Assume that π_ρ in (1.3) stands for a pdf associated to the variable $\boldsymbol{\theta}$. Let $p \geq 1$ such that $m_p^p := \int_{\mathbb{R}^d} \|\mathbf{u}\|^p K(\mathbf{u}) \mathrm{d}\mathbf{u} < \infty$. Then, for any $\rho > 0$, we have*

$$W_p(\pi, \pi_\rho) \leq \rho m_p. \tag{1.22}$$

Proof. See Appendix A.2. □

Note that (1.22) holds without assuming additional assumptions on the initial density π 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 π similarly to bias analysis in ABC (Sisson, Fan, and Beaumont, 2018a). Table 1.3 gives closed-form expressions of m_p when $p = 2$ for the multivariate generalizations of the kernels listed in Table 1.1. One can denote that the constant m_2 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 ρ rather than by the choice of K. In Section 1.4, we will illustrate Proposition 2 with numerical experiments.

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

Table 1.3: Closed-form expressions of m_2 appearing in (1.22) for multivariate generalizations of the kernels in Table 1.1.

1.3.2 *Pointwise bias for Bregman divergences*

In complement to Section 1.3.1 where κ_ρ was built using kernels, we now analyze the bias induced by considering π_ρ when κ_ρ is derived from a Bregman divergence d_ψ, see (1.8), that is

$$\kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) \propto \exp\left(-\frac{d_\psi(\mathbf{z}, \boldsymbol{\theta})}{\rho}\right). \tag{1.23}$$

Under infinite differentiability assumptions on both π and κ_ρ, we show that the pointwise bias $\pi_\rho - \pi$ is of the order of $\mathcal{O}(\rho)$ when ρ is sufficiently small, see Proposition 3.

Proposition 3. *Let $\mathrm{dom} f = \mathbb{R}^d$. Assume that π is analytic and twice differentiable on $\mathbb{R}^d$ and so does d_ψ 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 π 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*

Here $\|\mathbf{M}\|$ denotes the spectral norm of the matrix $\mathbf{M}$, i.e., its largest singular value.

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

it follows that

$$\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{\rho}). \tag{1.24}$$

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

$$\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}) = \frac{\rho}{2}\mathrm{Trace}\left(\mathbf{H}_\pi(\boldsymbol{\theta})\mathbf{H}_{d_\psi}(\boldsymbol{\theta})^{-1}\right) + o(\rho). \tag{1.25}$$

Proof. See Appendix A.3. □

Note that when $\psi(\mathbf{z}) = \|\mathbf{z}\|^2/2$, κ_ρ stands for a Gaussian smoothing kernel, see Section 1.3.1. In that case, we have the sanity check that the dependence w.r.t. ρ of the bias between π and π_ρ in (1.25) is the same as the one derived by Sisson, Fan, and Beaumont (2018a) when interpreting κ_ρ as a kernel.

1.3.3 *A detailed non-asymptotic analysis for Gaussian smoothing*

The previous sections gave quantitative approximation results for a large class of densities κ_ρ built either via a kernel or a Bregman divergence. In this section, we provide complementary results by restricting our analysis to the Gaussian smoothing case, that is

$$\kappa_\rho(\mathbf{z};\boldsymbol{\theta}) = \frac{e^{-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}}}{(2\pi\rho^2)^{d/2}}. \tag{1.26}$$

This particular yet convenient assumption will allow to complement and sharpen results of the two previous sections by deriving quantitative bounds which take into account the regularity properties of f. Furthermore, these bounds can be extended to composite potential functions $f = \sum_i f_i$ and used to assess the bias associated to 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, Liechty, and Müller, 2009; Dümbgen and Rufibach, 2009), and can be related to both the ADMM in optimization (Boyd et al., 2011; Vono, Dobigeon, and Chainais, 2019a) and the approximation involved in proximal MCMC methods, see the work by Pereyra (2016) and Section 5. Unfortunately, a straigthforward generalization of the proof techniques used in the sequel

Recall that f stands for the potential function of π that is $\pi \propto e^{-f}$.

does not yield informative upper bounds for smoothing associated to other Bregman divergences.

Lipschitz continuous 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 $\mathrm{dom} f = \mathbb{R}^d$ and such that there exists $L \geq 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^d$, $|f(\boldsymbol{\theta}) - f(\boldsymbol{\eta})| \leq L \|\boldsymbol{\theta} - \boldsymbol{\eta}\|$. Then,*

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq 1 - \Delta_d(\rho), \tag{1.27}$$

where

$$\Delta_d(\rho) = \frac{D_{-d}(L\rho)}{D_{-d}(-L\rho)}. \tag{1.28}$$

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

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz - x^2/2} x^{d-1} \mathrm{d}x. \tag{1.29}$$

Proof. See Appendix A.4. □

As expected from Corollary 1, note that this bound tends towards zero when $\rho \to 0$. 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 associated to f and the tolerance parameter ρ. In the limiting case $\rho \to 0$, the following equivalent function for the upper bound derived in (1.27) holds.

Corollary 2. *Let f such that Theorem 1 holds. In the limiting case $\rho \to 0$, we have:*

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \rho L \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} + o(\rho). \tag{1.30}$$

Proof. See Appendix A.5. □

Γ stands for the gamma function defined, for all $z > 0$, by $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} \mathrm{d}x$.

Using Stirling-like approximations when d is large in the equivalence relation (1.30) gives a mild dependence on the dimensionality of the problem in $\mathcal{O}\left(L\sqrt{d}\right)$. Potential functions verifying the hypothesis of Theorem 1 are common in machine learning and signal/image processing problems, see Section 1.4.3. 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 $L = \tau\sqrt{d}$ and satisfies Theorem 1 and Corollary 2. In this case, the dependence of (1.30) is linear w.r.t. d when d and ρ are sufficiently large and small, respectively.

Figure 1.5 gives the behavior of the upper bound in (1.27) w.r.t. the dimensionality d of the problem ranging from 1 to 10^6 and as a function of ρ in log-log scale. The linear relation between this upper bound and ρ shown in (1.30) is clearly observed for small values of ρ. Nonetheless, this

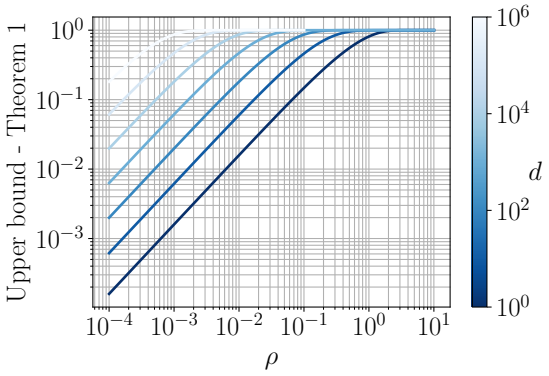


Figure 1.5: Behavior of the quantitative bound shown in Theorem 1 w.r.t. ρ in log-log scale for a set of dimensions d. The other quantities appearing in the bound have been set to 1.

upper bound is not a silver bullet. Indeed, as expected, for a fixed value of the parameter ρ, the approximation error increases as the dimension d grows. Thus, this bound suffers from the curse of dimensionality and become non-informative in high-dimension if ρ is not sufficiently small.

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 π and π_ρ. Similarly to the definition of the potential function f in (1.1), we define the potential function f_ρ associated to the approximate marginal π_ρ in (1.3), for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$f_\rho(\boldsymbol{\theta}) = -\log \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z})\right) \kappa_\rho(\mathbf{z};\boldsymbol{\theta}) \mathrm{d}\mathbf{z}. \tag{1.31}$$

By considering a Gaussian smoothing kernel κ_ρ, the potential f_ρ becomes

$$f_\rho(\boldsymbol{\theta}) = -\log \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2}\left\|\mathbf{z}-\boldsymbol{\theta}\right\|^2\right) \mathrm{d}\mathbf{z} + \frac{d}{2}\log(2\pi\rho^2). \tag{1.32}$$

Although the detailed study of f_ρ will be undertaken in Chapter 5, we already point out the following result which uniformly bounds the bias between f_ρ and f.

Proposition 4. *Let f satisfy the assumptions in Theorem 1. Then, for all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$L_\rho \leq f_\rho(\boldsymbol{\theta}) - f(\boldsymbol{\theta}) \leq U_\rho, \tag{1.33}$$

with

$$L_\rho = \log N_\rho - \log D_{-d}(-L\rho), \tag{1.34}$$
$$U_\rho = \log N_\rho - \log D_{-d}(L\rho), \tag{1.35}$$

and

$$N_\rho = \frac{2^{d/2-1}\Gamma\left(d/2\right)}{\Gamma(d)\exp\left(L^2\rho^2/4\right)}. \tag{1.36}$$

Proof. See Appendix A.6. □

When π 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 π_ρ stands for an approximation of the original target distribution π, it is important to control the credibility regions under π_ρ w.r.t. those drawn under π. 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 π_ρ and π, as stated below.

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

$$(1-\alpha)\frac{N_\rho}{D_{-d}(-L\rho)} \leq \int_{\mathcal{C}_\alpha^\rho} \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq \min\left(1, (1-\alpha)\frac{N_\rho}{D_{-d}(L\rho)}\right), \tag{1.37}$$

where $N_\rho = \dfrac{2^{d/2-1}\Gamma\left(d/2\right)}{\Gamma(d)\exp\left(L^2\rho^2/4\right)}$.

Proof. See Appendix A.7. □

Convex and smooth 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 $\mathrm{dom} f = \mathbb{R}^d$ and such that the following assumptions hold.*

(A_1) f is continuously differentiable and has an M-Lipschitz continuous gradient w.r.t. $\|\cdot\|$, that is $\exists\ M \geq 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^d$, $\left\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\eta})\right\| \leq M\left\|\boldsymbol{\theta} - \boldsymbol{\eta}\right\|$.

(A_2) f is convex, that is for every $\alpha \in [0,1]$, $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^d$, $f(\alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\eta}) \leq \alpha f(\boldsymbol{\theta}) + (1-\alpha)f(\boldsymbol{\eta})$.

Then, we have:

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \frac{d}{2}M\rho^2. \tag{1.38}$$

Proof. See Appendix A.8. □

Composite potential – Theorem 1 is easily extended to the case where the initial density π is expressed as a product of several terms which might involve linear operators acting on the variable of interest. If π stands for the pdf associated to the variable $\boldsymbol{\theta}$, this boils down to considering

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\sum_{i=1}^{b} f_i(\mathbf{A}_i\boldsymbol{\theta})\right), \tag{1.39}$$

where $b \geq 1$, for all $i \in [b]$, $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$, and a natural generalization of AXDA which writes

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b}) \propto \exp\left(-\sum_{i=1}^{b}\left[f_i(\mathbf{z}_i) + \frac{1}{2\rho^2}\left\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\right\|^2\right]\right). \tag{1.40}$$

In this scenario, a simple sufficient assumption to ensure that $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ is a probability density function is stated below, see also Proposition 6.

(A_0) For every $j \in [b]$, $\inf_{\mathbf{z}_j \in \mathbb{R}^{d_j}} f_j(\mathbf{z}_j) > -\infty$ (f_j are bounded from below), and for at least one $i \in [b]$ we have $d_i = d$, $\mathbf{A}_i$ is full rank, and $\exp(-f_i(\mathbf{z}_i))$ integrable on $\mathbb{R}^d$. (1.41)

Proposition 6 (Integrability of $\pi_\rho(\boldsymbol{\theta})$)**.** *Under Assumption* (A_0) *in* (1.41)*,* $\pi_\rho(\boldsymbol{\theta}) = \int \pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})\mathrm{d}\mathbf{z}_{1:b}$ *is integrable.*

Proof. First notice that using the conditional independence of $\mathbf{z}_j$ given $\boldsymbol{\theta}$ for all $j \in [b]$, we have

$$\pi_\rho(\boldsymbol{\theta}) \propto \prod_{j=1}^{b}\int_{\mathbf{z}_j \in \mathbb{R}^{d_j}} \exp\left(-f_j(\mathbf{z}_j) - \frac{\left\|\mathbf{z}_j - \mathbf{A}_j\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_j/2}}\mathrm{d}\mathbf{z}_j.$$

Now notice that for every $j \in [b]$, we have

$$\begin{aligned}
&\int_{\mathbf{z}_j \in \mathbb{R}^{d_j}} \exp\left(-f_j(\mathbf{z}_j) - \frac{\left\|\mathbf{z}_j - \mathbf{A}_j\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_j/2}}\mathrm{d}\mathbf{z}_j \\
&\leq \exp\left(-\inf_{\mathbf{z}_j} f_j(\mathbf{z}_j)\right)\int_{\mathbf{z}_j \in \mathbb{R}^{d_j}} \exp\left(-\frac{\left\|\mathbf{z}_j - \mathbf{A}_j\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_j/2}}\mathrm{d}\mathbf{z}_j \\
&= \exp\left(-\inf_{\mathbf{z}_j} f_j(\mathbf{z}_j)\right).
\end{aligned}$$

Using this bound for every $j \neq i$ (i as in Assumption (A_0)), we have

$$\begin{aligned}
&\prod_{j=1}^{b}\int_{\mathbf{z}_j \in \mathbb{R}^{d_j}} \exp\left(-f_j(\mathbf{z}_j) - \frac{\left\|\mathbf{z}_j - \mathbf{A}_j\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_j/2}}\mathrm{d}\mathbf{z}_j \\
&\leq C\int_{\mathbf{z}_i \in \mathbb{R}^{d_i}} \exp\left(-f_i(\mathbf{z}_i) - \frac{\left\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_i/2}}\mathrm{d}\mathbf{z}_i,
\end{aligned}$$

for some finite constant C. By integrating the integral term on the right hand side w.r.t. $\boldsymbol{\theta}$, we obtain

$$\begin{aligned}
&\int_{\boldsymbol{\theta}\in\mathbb{R}^d}\int_{\mathbf{z}_i \in \mathbb{R}^{d_i}} \exp\left(-f_i(\mathbf{z}_i) - \frac{\left\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_i/2}}\mathrm{d}\mathbf{z}_i\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\mathbf{z}_i \in \mathbb{R}^{d_i}}\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp\left(-f_i(\mathbf{z}_i) - \frac{\left\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\right\|^2}{2\rho^2}\right)\cdot\frac{1}{(2\pi\rho^2)^{d_i/2}}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{z}_i \\
&= \frac{\det\left(\mathbf{A}_i^T\mathbf{A}_i\right)^{-1/2}}{(2\pi\rho^2)^{d_i/2}}\int_{\mathbf{z}_i \in \mathbb{R}^{d_i}} \exp\left(-f_i(\mathbf{z}_i)\right)\int_{\mathbf{u}_i \in \mathbb{R}^{d_i}} \exp\left(-\frac{1}{2\rho^2}\left\|\mathbf{z}_i - \mathbf{u}_i\right\|^2\right)\mathrm{d}\mathbf{u}_i\mathrm{d}\mathbf{z}_i \\
&= \det\left(\mathbf{A}_i^T\mathbf{A}_i\right)^{-1/2}\cdot\int_{\mathbf{z}_i \in \mathbb{R}^{d_i}} \exp\left(-f_i(\mathbf{z}_i)\right)\mathrm{d}\mathbf{z}_i,
\end{aligned}$$

which is finite using the integrability condition on $\mathbf{z}_i$. Hence $\pi_\rho(\boldsymbol{\theta})$ is integrable. □

Under this assumption, we have the following corollary.

Corollary 3. *Assume* (A_0) *in* (1.41). *For all* $i \in [b]$, *let* f_i *satisfy the assumptions of Theorem 1. Then,*

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq 1 - \prod_{i=1}^{b} \Delta_d^{(i)}(\rho), \tag{1.42}$$

where $\Delta_d^{(i)}(\rho) = D_{-d_i}(L_i\rho)/D_{-d_i}(-L_i\rho)$ *and* L_i *is the Lipschitz constant associated to the function* f_i.

Proof. See Appendix A.9. □

In the case where the potential f is not convex and smooth, we do not have for the moment an informative and quantitative result.

Smooth and strongly convex potential – The following proposition builds on the heat equation to derive an explicit and simple bound on the bias between π_ρ and π in 1-Wasserstein distance when the potential f is sufficiently smooth and strongly convex.

Proposition 7. *Let* f *be convex, twice differentiable,* M*-Lipschitz continuous and such that* $f(\boldsymbol{\theta}) \geq a_1 + a_2\|\boldsymbol{\theta}\|^\alpha$ *for some* $a_1 \in \mathbb{R}$, $a_2 > 0$ *and* $\alpha > 0$. *Then, we have:*

$$W_1(\pi, \pi_\rho) \leq \min\left(\rho\sqrt{d}, \frac{1}{2}\rho^2\sqrt{Md}\right). \tag{1.43}$$

When f is m-strongly convex, note that this growth condition is satisfied with $a_1 = m\left\|\boldsymbol{\theta}^\star\right\|^2/2$, $a_2 = m/2$ and $\alpha = 2$, where $\boldsymbol{\theta}^\star$ is the minimum of f.

Proof. See Appendix A.10. □

Note that (1.43) sharpens the general result shown in Proposition 2.

Indicator function of a convex body – All the previous results assumed that $\mathrm{dom} f = \mathbb{R}^d$ and already cover a large class of potential functions and associated density functions used in practice. We complement these results by now focusing on densities with bounded support, which commonly appear in statistical machine learning and signal processing when one wants to estimate parameters subject to contraints on the parameter space (Klein and Moeschberger, 2005; Johnson and Albert, 2006; Celeux et al., 2012; Paisley, Blei, and Jordan, 2014). Such a bounded support might also yield truncated densities from which it is difficult to sample (Betancourt, 2011; Altmann, McLaughlin, and Dobigeon, 2014). More precisely, we consider here a convex body $\mathcal{K} \subset \mathbb{R}^d$, i.e., a compact convex set with non-empty interior, and a potential function $f \coloneqq \iota_{\mathcal{K}}$ standing for the indicator function of $\mathcal{K}$ and defined for $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$\iota_{\mathcal{K}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} \in \mathcal{K}, \\ +\infty & \text{if } \boldsymbol{\theta} \notin \mathcal{K}. \end{cases} \tag{1.44}$$

In order to quantify the bias between π_ρ and π in this case, we will build on the recent work by Brosse et al. (2017) where the authors analyzed the bias between π and an approximate density $\tilde{\pi}_\rho$ defined, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$\tilde{\pi}_\rho(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathrm{proj}_{\mathcal{K}}(\boldsymbol{\theta})\right\|^2\right), \tag{1.45}$$

Note that the potential function of $\tilde{\pi}_\rho$ stands for the Moreau envelope of $\iota_{\mathcal{K}}$ (Moreau, 1965).

with $\mathrm{proj}_{\mathcal{K}}(\boldsymbol{\theta})$ the projection of $\boldsymbol{\theta}$ onto $\mathcal{K}$. In Chapter 5, we will see that the potential function of $\tilde{\pi}_\rho$ is strongly related to that of π_ρ. Combining this result with the work by Brosse et al. (2017), the following bounds hold.

Proposition 8. *Let $f = \iota_{\mathcal{K}}$ with $\mathcal{K}$ a convex body containing the origin. Assume that there exists $r > 0$ such that $\mathcal{B}(\mathbf{0}_d, r) \subset \mathcal{K}$. Then, for all $\rho > 0$,*

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \sum_{i=1}^{d} \left(\frac{\sqrt{2}\rho d}{r}\right)^i. \tag{1.46}$$

In addition if $\rho \in \left(0, \frac{r}{2\sqrt{2}d}\right]$, we have:

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \frac{2\sqrt{2}\rho d}{r}. \tag{1.47}$$

Proof. See Appendix A.11. □

The set $\mathcal{B}(\mathbf{0}_d, r)$ stands for the closed ball of center $\mathbf{0}_d$ and radius r, that is $\mathcal{B}(\mathbf{0}_d, r) = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\| \leq r\}$.

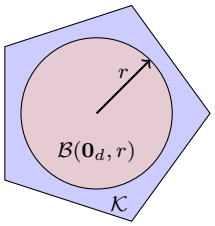


Figure 1.6: Illustration of the assumption $\mathcal{B}(\mathbf{0}_d, r) \subset \mathcal{K}$ used in Proposition 8.

From the last bound in (1.47) and similarly to the previous non-asymptotic results, one can denote that the choice of ρ depends on the dimension d. More precisely, to achieve a prescribed precision ϵ its choice is inversely proportional to the dimension d. In this scenario, one can note in addition that ρ has to be chosen such that it is proportional to the radius of the ball $\mathcal{B}(\mathbf{0}_d, r)$.

1.3.4 *Summary*

Table 1.4 recaps the theoretical results shown in Section 1.3 and highlights scalings in ρ and d in our bounds. First, we highlight asymptotic results associated to the two general approaches to build the density κ_ρ, namely via standard kernels or by using divergence functions. For divergence functions, note that we do not have access to the dependence w.r.t. the dimension because the latter is hidden in the Trace term in Proposition 3. Then, we focus on the Gaussian smoothing case which stands for a special instance of the two previous general approaches. This specific case allows to complement the general results depending on the regularity properties of the potential function f (e.g., convexity or Lipschitz continuity). Overall, this table allows to understand and promote the proposed approximation in a large variety of scenarios.

1.4 Numerical illustrations

This section illustrates the proposed approximation and the quantitative results shown in Section 1.3. As shown in Table 1.3, the bias induced by considering π_ρ is mostly driven by the value of the tolerance parameter ρ rather than by the choice of κ_ρ. Hence, for simplicity, most of the numerical illustrations hereafter consider the case where κ_ρ is a Gaussian smoothing kernel.

$\kappa_\rho(\mathbf{z};\boldsymbol{\theta})$	conditions on f	regime	quantity of interest	ρ	d
$\rho^{-d}K\left(\frac{\boldsymbol{\theta}-\mathbf{z}}{\rho}\right)$	$\mathrm{dom} f = \mathbb{R}^d$	$\rho > 0$	$W_1(\pi, \pi_\rho)$	$\mathcal{O}(\rho)$	$\mathcal{O}(\sqrt{d})$
$\exp\left(-\frac{d_\psi(\mathbf{z},\boldsymbol{\theta})}{\rho}\right)$	$\mathrm{dom} f = \mathbb{R}^d$, $\mathcal{C}^\infty(\mathbb{R}^d)$	$\rho \to 0$	$\pi_\rho - \pi$	$\mathcal{O}(\sqrt{\rho})$	-
$\frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{2\pi\rho^2}$	$\mathrm{dom} f = \mathbb{R}^d$, coercive, convex & smooth	$\rho > 0$	$W_1(\pi, \pi_\rho)$	$\mathcal{O}(\rho^2)$	$\mathcal{O}(\sqrt{d})$
	$\mathrm{dom} f = \mathbb{R}^d$, Lipschitz	$\rho \to 0$, $d \to +\infty$	$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}}$	$\mathcal{O}(\rho)$	$\mathcal{O}(\sqrt{d})$
	$\mathrm{dom} f = \mathbb{R}^d$, coercive, convex & smooth	$\rho \to 0$	$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}}$	$\mathcal{O}(\rho^2)$	$\mathcal{O}(d)$
	$f = \iota_{\mathcal{K}}$, $\mathcal{B}(\mathbf{0}_d, r) \subset \mathcal{K}$	$0 < \rho \leq \frac{r}{2\sqrt{2}d}$	$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}}$	$\mathcal{O}(\rho)$	$\mathcal{O}(d)$

Table 1.4: Summary of the orders of magnitude of ρ and d based on the quantitative results shown in Section 1.3. The notation "-" in the last column of the second line means that the scaling is not explicitly available.

1.4.1 *Multivariate Gaussian example*

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

$$\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.48}$$

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

$$\pi_\rho(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma} + \rho^2\mathbf{I}_d). \tag{1.49}$$

In particular, let consider the univariate setting. In this case, the variance under π_ρ is $\sigma^2 + \rho^2$ and simply corresponds to the variance under π inflated by a factor ρ^2. Therefore, the approximation will be reasonable if ρ^2/σ^2 is sufficiently small, see Figure 1.7. In this figure, we also show the approximation induced by considering a uniform kernel (see Table 1.2) instead of a Gaussian one. The smoothing via the uniform kernel performs slightly better than Gaussian smoothing due to its lower variance ($\rho^2/3$ instead of ρ^2 for the Gaussian kernel). In both cases, the approximation is reasonable for small ρ although π_ρ, built with a uniform kernel, no longer belongs to the Gaussian family.

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

$$\Sigma_{ij} = 2\exp\left(-\frac{(s_i - s_j)^2}{2a^2}\right) + 10^{-6}\delta_{ij}, \forall i, j \in [d] \tag{1.50}$$

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 1.8 shows the behavior of the quantitative bounds derived in Proposition 2, Proposition 7 and Theorem 2. The Gaussian case allows

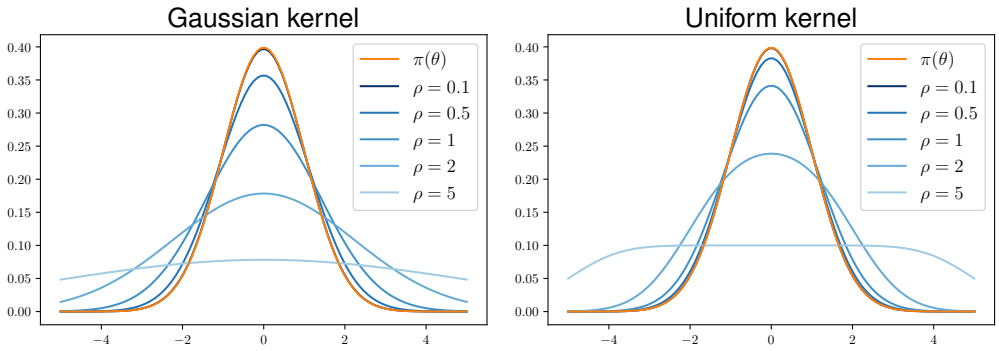


Figure 1.7: Bias between π_ρ and π in the case $\pi = \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma = 1$. (left) π_ρ is built with a Gaussian kernel $\mathcal{N}(0, \rho^2)$ and (right) with a uniform kernel on $[-\rho, \rho]$. Note that the curves associated to π and π_ρ for $\rho = 0.1$ are overlapping.

to compute exactly $W_2(\pi, \pi_\rho)$ for $d \geq 1$ by noting that $W_2^2(\pi, \pi_\rho) = \mathrm{Trace}(\boldsymbol{\Sigma} + \rho^2 \mathbf{I}_d - 2\rho\boldsymbol{\Sigma}^{1/2})$. The 1-Wasserstein distance appearing in Proposition 7 admits a simple expression in the univariate setting, that is $W_1(\pi, \pi_\rho) = \int_{\mathbb{R}} |F(u) - F_\rho(u)| \mathrm{d}u$, where F and F_ρ are the cumulative distribution functions associated to π and π_ρ, respectively. Finally, $\|\pi - \pi_\rho\|_{\mathrm{TV}}$ 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 ρ since it does not catch the behavior in $\mathcal{O}(\rho^2)$ when ρ 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 π such as differentiability or strong convexity of its potential. On the contrary, the bounds on the 1-Wasserstein and total variation distances, derived under stronger assumptions, manage to achieve the correct rate of the order $\mathcal{O}(\rho^2)$ for small ρ.

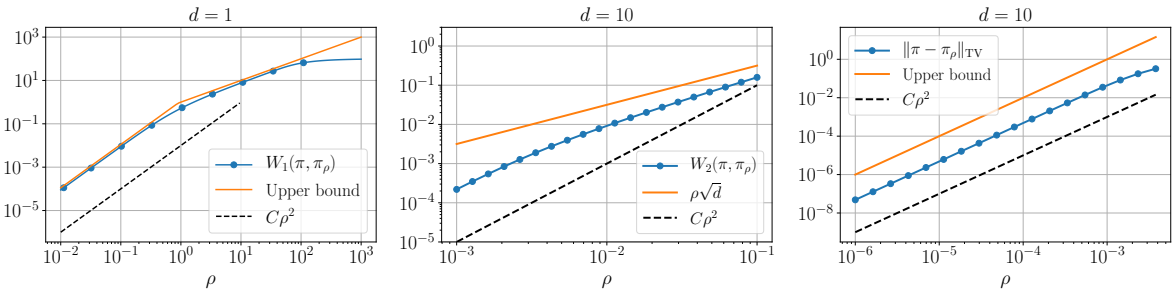


Figure 1.8: From left to right: Illustration of the quantitative bounds (1.43), (1.22) and (1.38) associated to 1-Wasserstein, 2-Wasserstein and total variation distances, respectively. The decay in $\mathcal{O}(\rho^2)$ is shown via the dashed line $C\rho^2$ where C is a constant.

1.4.2 *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 $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\pi(\boldsymbol{\theta}) \coloneqq \pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 - g(\mathbf{B}\boldsymbol{\theta})\right), \qquad (1.51)$$

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

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, with a Gaussian choice for κ_ρ, the joint density π_ρ writes

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 - g(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{B}\mathbf{x} - \mathbf{z}\|^2\right). \quad (1.52)$$

Then, a Gibbs algorithm can be used to sample from this joint probability distribution. Since Chapter 2 is dedicated to this sampler, we do not detail its associated sampling steps here. We rather highlight only the main tools which permit an efficient sampling from (1.52). For the $\mathbf{z}$-conditional, one can for instance use a simple data augmentation scheme (Park and Casella, 2008). 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, see for instance the works by Papandreou and Yuille (2011), Barbos et al. (2017), and Marnissi et al. (2018) and Chapter 4.

In this specific case, the potential g_ρ associated to the smoothed prior distribution, see (1.32), has a closed-form expression given for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$\begin{aligned} g_\rho(\boldsymbol{\theta}) &= \frac{k}{2}\log(2\pi\rho^2) - \log\prod_{i=1}^{k}\int_{\mathbb{R}}\exp\left(-\tau|z_i| - \frac{1}{2\rho^2}(\mathbf{b}_i^T\boldsymbol{\theta} - z_i)^2\right)\mathrm{d}z_i \\ &= \frac{k}{2}\log(2\pi\rho^2) - \log\prod_{i=1}^{k}\left(a(\boldsymbol{\theta})\left[\exp\left(b(\boldsymbol{\theta})^2\right)\left\{1 - \mathrm{erf}(b(\boldsymbol{\theta}))\right\} + \exp\left(c(\boldsymbol{\theta})^2\right)\left\{1 - \mathrm{erf}(c(\boldsymbol{\theta}))\right\}\right]\right), \end{aligned} \quad (1.53)$$

with $a(\boldsymbol{\theta}) = \sqrt{\pi\rho^2/2}\exp\left(-(\mathbf{b}_i^T\boldsymbol{\theta})^2/(2\rho^2)\right)$, $b(\boldsymbol{\theta}) = \sqrt{\rho^2/2}(\tau - \mathbf{b}_i^T\boldsymbol{\theta}/\rho^2)$, $c(\boldsymbol{\theta}) = \sqrt{\rho^2/2}(\tau + \mathbf{b}_i^T\boldsymbol{\theta}/\rho^2)$ 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_ρ has no closed form, one can estimate it by a Monte Carlo approximation. Figure 1.9 shows the behavior of the regularized potential g_ρ defined in (1.53) for several values of the parameter ρ 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 τ has been set to $\tau = 1$. The contours of the shaded area correspond to $g + L_\rho$ and $g + U_\rho$. The potential g_ρ is a smooth approximation of the potential g associated to the initial prior as expected, see Property *(iv)* in Proposition 1. Note that the inequalities derived in (1.33) are verified. The third row of Figure 1.9 shows the form of the posterior of θ_1 defined in (1.52) for $y = 1$, $x = 2$ and $\sigma = 1$ and

derived from the smoothed prior distributions shown in Figure 1.9. For sufficiently small values of ρ, the marginal π_ρ stands for a quite accurate approximation of the original target π.

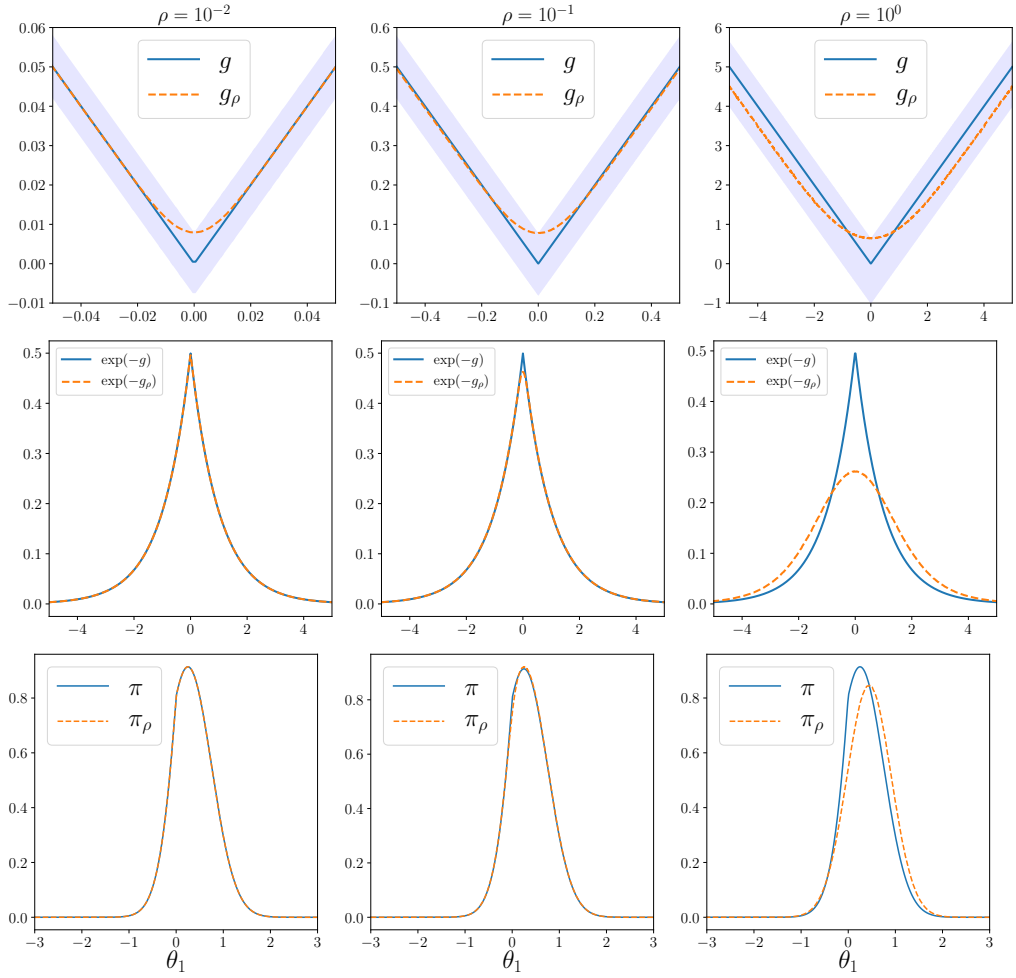


Figure 1.9: From left to right, $\rho = 0.01$, $\rho = 0.1$ and $\rho = 1$. (1st row) Behaviors of g (blue) and g_ρ (orange) where the contours of the shaded area correspond to $g + L_\rho$ and $g + U_\rho$; (2nd row) the corresponding normalized smoothed prior densities proportional to $\exp(-g)$ and $\exp(-g_\rho)$; (3rd row) posterior densities π_ρ w.r.t. ρ.

Table 1.5 illustrates the bounds derived in (1.37) for several values of ρ. For each case, the values of the bounds are summarized in the interval

$$\mathcal{I}_\alpha^\rho = [(1-\alpha)N_\rho / D_{-d}(-L\rho), \min(1, (1-\alpha)N_\rho / D_{-d}(L\rho)], \quad (1.54)$$

and the real coverage $\int_{\mathcal{C}_\alpha^\rho} \pi(\theta_1)\mathrm{d}\theta_1$ is also reported. The (1-α)-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 ρ is sufficiently small which is not surprising since the assumptions on the potential of π_ρ 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 $\int_{\mathcal{C}_\alpha^\rho} \pi(\theta_1)\mathrm{d}\theta_1$, we emphasize that the marginal π_ρ stands for a conservative approximation of π in this example. Indeed, in each case, the (1-α)-credibility interval under π_ρ denoted $\mathcal{C}_\alpha^\rho$ covers at least $100(1-\alpha)\%$ of the probability mass under π.

ρ	$\mathcal{C}_\alpha$	$\mathcal{C}_\alpha^\rho$	$\int_{\mathcal{C}_\alpha^\rho} \pi(\theta_1)\mathrm{d}\theta_1$	$\mathcal{I}_\alpha^\rho$
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]

Table 1.5: Illustration of the bound derived in (1.37) for the marginal posterior π_ρ depicted in Section 1.4.2. The (1-α)-credibility intervals $\mathcal{C}_\alpha$ and $\mathcal{C}_\alpha^\rho$ are the highest posterior density regions associated to each density with $\alpha = 0.05$.

1.4.3 *Illustration for Lipschitz loss functions used in statistical learning*

Some of the results of Section 1.3.3 assume that the potential function f associated to π is Lipschitz. Interestingly, such Lipschitz functions are used in standard statistical learning problems to evaluate the discrepancy between observations and model outputs (van de Geer, 2016). Table 1.6 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 1.10 illustrates the form of these losses and associated regularized potentials f_ρ 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 (1.39) with

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

Table 1.6: 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.

$$f_j(y_j;\boldsymbol{\theta}) = f(y_j;\mathbf{x}_j^T\boldsymbol{\theta}), \tag{1.55}$$

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 1.6 and $\boldsymbol{\theta}\in\mathbb{R}^d$ is the parameter to infer. Since all the loss functions listed in Table 1.6 are Lipschitz continuous w.r.t. their second argument t with Lipschitz constant equal to 1, the potential f_j in (1.55) is also Lipschitz with constant $L_j = \|\mathbf{x}_j\|$. Motivated by the robustness properties inherited by AXDA, see Section 1.2.3, we consider the smoothing of the likelihood contribution f_j associated to each observation with a Gaussian kernel. The results of Corollary 3 can then be applied to π defined in (1.39).

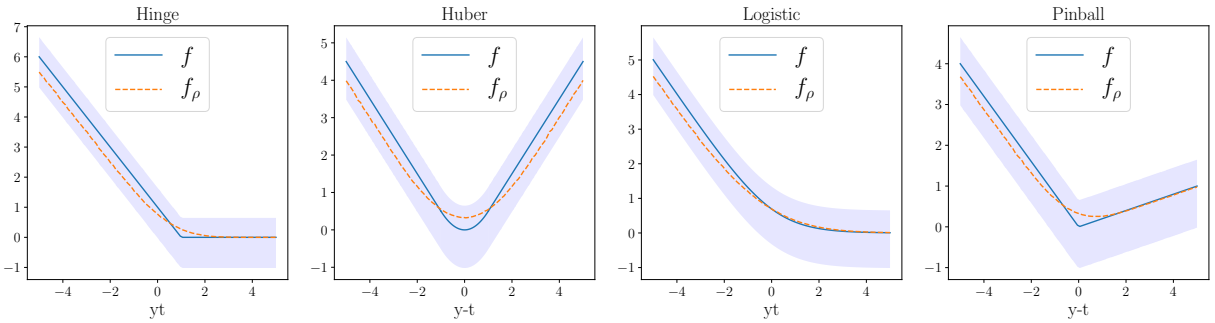


Figure 1.10: Loss functions of Table 1.6 along with their associated regularized loss f_ρ with $\rho = 1$ estimated with a Monte Carlo approximation. The Huber and pinball losses have been plotted with $\delta = 1$ and $\tau = 0.2$, respectively. The contours of the shaded area correspond to $f + L_\rho$ and $f + U_\rho$.

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 ranging 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 $\mathbf{X}^T = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ 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 1.11 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 ρ as the size n of the dataset increases. Nonetheless, the effect of n on the bound is not highly prohibitive. In the two cases $d = 10$ and $d = 10^3$, ρ 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 ρ is small is of the order $\mathcal{O}(n\rho)$ for a fixed dimension d. 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. (2018).

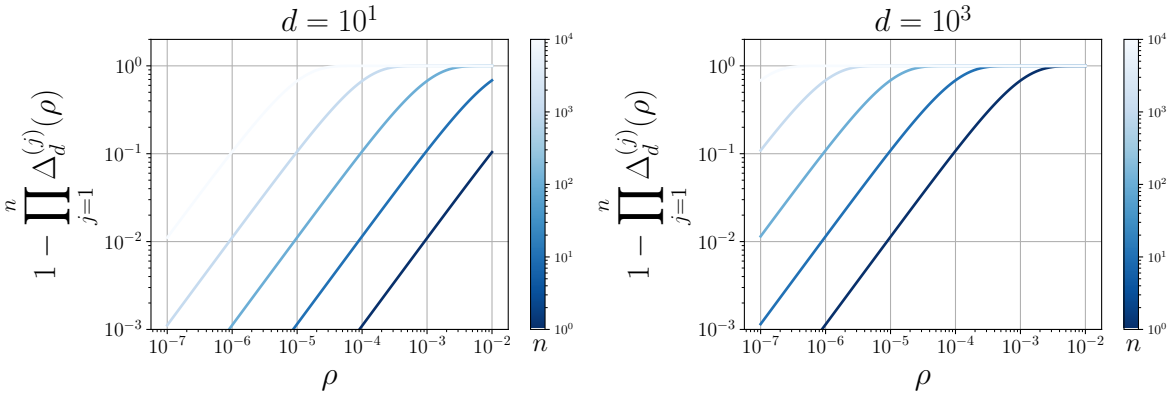


Figure 1.11: Behavior of the upper bound in Corollary 3 w.r.t. ρ and n for several values of the dimension d. The notation $\Delta_d^{(j)}(\rho)$ has been defined in Corollary 3.

1.5 Conclusion

This chapter presented a unifying framework for asymptotically exact data augmentation (AXDA) schemes. This framework introduced approximate densities built with a weak converging sequence in order to simplify the inference. By building on existing works which considered special instances of AXDA, we illustrated 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 encompassed a large class of AXDA models and a detailed non-asymptotic analysis has been conducted 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. The construction and analysis of this framework has been submitted to

an international journal (Vono, Dobigeon, and Chainais, 2020a).

The next chapters will demonstrate that AXDA models can remarkably improve the inference task in big data and high-dimensional settings. More precisely, Chapter 2 presents an efficient Gibbs sampler to sample from specific instances of AXDA models while Chapter 3 provides explicit mixing time bounds for this algorithm. In summary, at the price of an approximation which comes with theoretical guarantees, AXDA approaches will appear to be a general, systematic and efficient way to conduct 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.

Monte Carlo sampling from AXDA

2

"Divide ut regnes."

— used by Julius Ceasar

Chapter contents

The previous chapter presented the general AXDA framework and its associated approximate models which can be used to perform likelihood or posterior inferences. Within the Bayesian paradigm, we now consider the specific problem of sampling from a given posterior distribution. As in Chapter 1, we assume that sampling exactly from this posterior distribution is difficult and propose to rely on AXDA to simplify the sampling task.

This chapter focuses on a specific Monte Carlo sampling algorithm to perform such an approximate posterior inference. More precisely, a Gibbs sampler (Geman and Geman, 1984) will be presented and applied to several high-dimensional Bayesian inference problems demonstrating the benefits of the proposed AXDA framework. Interestingly, we will show that this sampler shares strong connections with quadratic penalty methods in optimization. Hence, similarly to previous works (Duane et al., 1987; Roberts and Tweedie, 1996; Pereyra, 2016), this chapter contributes to fill the gap between simulation and optimization by drawing new connections between these two fields. These connections will be strengthened in Chapter 5.

The so-called *split Gibbs sampler* (SGS) is introduced in Section 2.1. In Section 2.2, we instantiate this MCMC algorithm on three Bayesian inference problems which are classically encountered in statistical signal processing and machine learning. Experiments associated to these problems are then presented in Section 2.3.

The major part of the results of this chapter has been published in an international journal and has been presented at international and national conferences:

M. Vono, N. Dobigeon, and P. Chainais (2019a). "Split-and-augmented Gibbs sampler - Application to large-scale inference problems." *IEEE Transactions on Signal Processing* 67 (6): 1648–1661. doi:10.1109/TSP.2019.2894825

M. Vono, N. Dobigeon, and P. Chainais (2018). "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler." In *IEEE International Workshop on Machine Learning for Signal Processing*. doi:10.1109/MLSP.2018.8516963

M. Vono, N. Dobigeon, and P. Chainais (2019c). "Efficient sampling through variable splitting-inspired Bayesian hierarchical models." In *IEEE*

International Conference on Acoustics, Speech, and Signal Processing. doi:10.1109/ICASSP.2019.8682982

M. Vono, N. Dobigeon, and P. Chainais (2019b). "Bayesian image restoration under Poisson noise and log-concave prior." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/ICASSP.2019.8683031

M. Vono, N. Dobigeon, and P. Chainais (2019e). "Un modèle augmenté asymptotiquement exact pour la restauration bayésienne d'images dégradées par un bruit de Poisson." In *Proc. of GRETSI*

2.1 Gibbs sampler

We consider the situation where one is interested in carrying out Bayesian inference about a parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ based on observed data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i$ are covariates associated to observation y_i. We define the observation set as $\mathcal{Y} := \{y_1, \ldots, y_n\}$ and its vectorized counterpart as $\mathbf{y} = [y_1, \ldots, y_n]^T$. Based on the AXDA framework introduced in Chapter 1, this section presents a MCMC algorithm to sample approximately from a target posterior distribution with pdf, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, which writes

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\sum_{i=1}^{b} f_i(\mathbf{A}_i\boldsymbol{\theta})\right), \tag{2.1}$$

where for all $i \in [b]$, $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ and $\mathbf{A}_i \in \mathbb{R}^{d_i \times d}$ are such that π defines a proper pdf. The potentials $(f_i; i \in [b])$ are assumed to be possibly functions of all or a subset of the observations $\mathcal{Y}$, hereafter generically denoted $\boldsymbol{y}_i$ where $\boldsymbol{y}_i \subseteq \mathcal{Y}$ and $\bigcup_{i=1}^{n} \boldsymbol{y}_i = \mathcal{Y}$. Note that potentially we may have $\boldsymbol{y}_i = \emptyset$. In this case the corresponding potentials f_i are associated with a prior distribution assigned to $\boldsymbol{\theta}$. When $\boldsymbol{y}_i = \mathcal{Y}$, then a unique potential is associated to the likelihood function and the others are associated to the prior assigned to $\boldsymbol{\theta}$. To simplify notation, this dependence is notationally omitted.

2.1.1 *Split Gibbs sampler*

The composite potential function of the target density π in (2.1) stands for a sum of b potentials which involve linear operators $(\mathbf{A}_i; i \in [b])$ acting on the parameter to infer $\boldsymbol{\theta}$. Such a scenario for instance appears in statistical problems involving generalized (non-)linear models (see Section 1.2.1) and/or non-conjugate and non-differentiable prior distributions (Dupé, Fadili, and Starck, 2009) (see also Section 2.2.3). Due to this composite structure but also to a possible distributed architecture (see Section 1.2.2), sampling from π is challenging. To overcome these issues and ease posterior sampling, we propose to rely on the proposed AXDA framework. Among the b individual potential functions $(f_i; i \in [b])$, we can assume without loss of generality that only $p \in [b]$ of them raise sampling difficulties while the remaining $b - p$ potentials are supposed to admit a nice structure (e.g., quadratic and isotropic potential functions).

As such, we apply the approximation described in Section 1.1 p times to decouple the p "problematic" potentials. For the sake of simplicity, we consider here an isotropic Gaussian choice for κ_ρ, that is $\kappa_\rho(\mathbf{z}_i; \mathbf{A}_i\boldsymbol{\theta}) = \phi(\mathbf{z}_i; \mathbf{A}_i\boldsymbol{\theta}; \rho^2\mathbf{I}_{d_i})$ for $i \in [p]$. We could have considered an alternative prior for $\mathbf{z}_i$ (Dai Pra, Scoppola, and Scoppola, 2012; Rendell et al., 2018) as described in Chapter 1 but this choice is also motivated by the fact that the corresponding quadratic potential enjoys attractive properties such as smoothness and strong convexity. This yields an approximate joint posterior distribution π_ρ which admits the pdf

Here ϕ denotes the pdf associated to the Gaussian distribution.

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:p}) \propto \exp\left(-\sum_{i=p+1}^{b} f_i(\mathbf{A}_i\boldsymbol{\theta})\right) \prod_{i=1}^{p} \exp\left(-f_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right). \tag{2.2}$$

The main benefit of working with the joint distribution $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:p})$ defined by (2.2) instead of $\pi(\boldsymbol{\theta})$ is the fact that, under π_ρ, the conditional distribution of the auxiliary variables $\mathbf{z}_{1:p}$ given $\boldsymbol{\theta}$ factorizes across $i \in [p]$, i.e., $\pi_\rho(\mathbf{z}_{1:p}|\boldsymbol{\theta}) = \prod_{i=1}^{p} \pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ where

$$\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}) \propto \exp\left(-f_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right). \tag{2.3}$$

Hence these simulation steps can be performed in parallel and are expected to be simpler than sampling directly from (2.1). Moreover, the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{z}_{1:p}$ writes

$$\pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:p}) \propto \exp\left(-\sum_{i=p+1}^{b} f_i(\mathbf{A}_i\boldsymbol{\theta}) - \sum_{i=1}^{p} \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right). \tag{2.4}$$

Note that if the potentials $(f_i; p+1 \leq i \leq b)$ are quadratic, this conditional is Gaussian and can be efficiently sampled using techniques reviewed in Chapter 4.

This suggests using a Gibbs sampler to sample from $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:p})$. The resulting so-called split Gibbs sampler (SGS) is described in Algorithm 1. The adjective "split" is related to the variable splitting technique commonly used in numerical analysis to simplify computations, see Section 2.1.2. In scenarios when the conditional distributions $\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ cannot be sampled from exactly, we can also sample from this extended target using a time-discretized Langevin diffusion. This algorithm also allows us to update the auxiliary variables $\mathbf{z}_i$ in parallel; see Appendix B.1. The π-irreducibility and aperiodicity of SGS follows because SGS, defined on the extended state space including $\mathbf{z}_{1:p}$, is a Gibbs sampler with systematic scan, and it satisfies the positivity condition of Gibbs sampling (since the densities are always positive); see e.g., Roberts and Smith (1994).

2.1.2 *Connections with optimization approaches*

The SGS whose main steps are described in Algorithm 1 can be related to common optimization approaches. More precisely, it can be seen as the stochastic counterpart of alternating minimization (AM) algorithms based

Algorithm 1: Split Gibbs Sampler (SGS)

Input: Potentials f_i for $i \in [b]$, penalty parameter ρ, initialization $\boldsymbol{\theta}^{(0)}$ and nb. of iterations T.

for $t \leftarrow 1$ **to** T **do**
  for $i \leftarrow 1$ **to** p **do**
    $\mathbf{z}_i^{(t)} \sim \pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}^{(t-1)})$ (see Equation (2.3))
  end
  $\boldsymbol{\theta}^{(t)} \sim \pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:p}^{(t)})$ (see Equation (2.4))
end

on the classical quadratic penalty method (Nocedal and Wright, 2006, Chapter 7). Instead of minimizing a given composite objective function, these algorithms transform this unconstained minimization problem into a constrained one via a so-called variable splitting technique. This constraint is then relaxed by adding a "seemingly naive" quadratic term to the initial objective function before performing alternating minimization. In the sequel, we detail such an optimization approach and draw connections between the latter and Algorithm 1.

Quadratic penalty method – We consider the maximum a posteriori estimation problem under the posterior distribution π in (2.1), that is

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^{b} f_i(\mathbf{A}_i \boldsymbol{\theta}). \tag{2.5}$$

Similarly to direct sampling from π, solving directly this minimization problem might be computationally demanding because of a sum of b composite terms, the presence of linear operators acting on $\boldsymbol{\theta}$, non-differentiability or a possible distributed architecture. To bypass these issues, some authors (Wang et al., 2008; Afonso, Bioucas-Dias, and Figueiredo, 2010; van Leeuwen and Herrmann, 2015) proposed to build on variable splitting: they introduce a set of auxiliary variables $(\mathbf{z}_i; i \in [p])$ to reformulate (2.5) into the constrained minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \mathbb{R}^d, \mathbf{z}_1 \in \mathbb{R}^{d_1}, \ldots, \mathbf{z}_p \in \mathbb{R}^{d_p}} \sum_{i=p+1}^{b} f_i(\mathbf{A}_i \boldsymbol{\theta}) + \sum_{i=1}^{p} f_i(\mathbf{z}_i) \\ &\text{subject to} \quad \mathbf{z}_i = \mathbf{A}_i \boldsymbol{\theta}, i \in [p], \end{aligned} \tag{2.6}$$

where it is assumed that the functions $(f_i; p+1 \leq i \leq b)$ admit a simple structure and do not need any splitting. The constraint $\mathbf{z}_i = \mathbf{A}_i \boldsymbol{\theta}$ is then relaxed by adding a quadratic penalty term in the objective function. This yields the approximate joint minimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d, \mathbf{z}_1 \in \mathbb{R}^{d_1}, \ldots, \mathbf{z}_p \in \mathbb{R}^{d_p}} L(\boldsymbol{\theta}, \mathbf{z}_{1:p}) \coloneqq \sum_{i=p+1}^{b} f_i(\mathbf{A}_i \boldsymbol{\theta}) + \sum_{i=1}^{p} f_i(\mathbf{z}_i) + \frac{\|\mathbf{z}_i - \mathbf{A}_i \boldsymbol{\theta}\|^2}{2\rho^2}. \tag{2.7}$$

This optimization problem can be solved by alternating minimization (Beck,

2015). For fixed $\boldsymbol{\theta} \coloneqq \boldsymbol{\theta}^{(t-1)}$, one first minimizes $L(\boldsymbol{\theta}, \mathbf{z}_{1:p})$ w.r.t. $\mathbf{z}_i$ for each factor $i \in [p]$ before minimizing, for fixed $\mathbf{z}_{1:p} \coloneqq \mathbf{z}_{1:p}^{(t)}$, $L(\boldsymbol{\theta}, \mathbf{z}_{1:p})$ w.r.t. $\boldsymbol{\theta}$, see Algorithm 2.

Algorithm 2: Quadratic penalty method

Input: Potentials f_i for $i \in [b]$, penalty parameter ρ, initialization $\boldsymbol{\theta}^{(0)}$ and nb. of iterations T.

for $t \leftarrow 1$ **to** T **do**
  for $i \leftarrow 1$ **to** p **do**
    $\mathbf{z}_i^{(t)} \in \arg\min_{\mathbf{z}_i} -\log \pi_\rho(\mathbf{z}_i | \boldsymbol{\theta}^{(t-1)})$ (see Equation (2.3))
  end
  $\boldsymbol{\theta}^{(t)} \in \arg\min_{\boldsymbol{\theta}} -\log \pi_\rho(\boldsymbol{\theta} | \mathbf{z}_{1:p}^{(t)})$ (see Equation (2.4))
end

Similarly to SGS and at the price of an approximation, the main benefit of this approach is that the minimization problems w.r.t. each auxiliary variable now involves a single function f_i without any operator and a quadratic term. This suggests the use of proximal methods if the proximity operator (Moreau, 1965) of f_i is available in closed-form or can be easily approximated (Combettes and Pesquet, 2011).

SGS and quadratic penalty methods – Interestingly, these AM steps stand for the deterministic counterpart of the conditional sampling steps in Algorithm 1. Indeed, instead of drawing a random variable following each conditional, these minimization steps only find the mode associated to each conditional probability distribution and can be related to iterated conditional modes in image processing (Besag, 1986). This shows another interesting bond between optimization and simulation and complements earlier connections between these two fields. For instance, we can mention the celebrated one-to-one equivalence between gradient descent and discretized Langevin dynamics (Roberts and Tweedie, 1996; Pereyra, 2016; Durmus, Moulines, and Pereyra, 2018; Brosse, Moulines, and Durmus, 2018) and more recently the use of Hamiltonian dynamics to define first-order descent schemes achieving linear convergence (Duane et al., 1987; Maddison et al., 2018).

2.2 Application to Bayesian inference problems

In this section, we instantiate SGS on three challenging Bayesian inference problems which commonly appear in image processing namely (i) unsupervised image deconvolution with a smooth prior, (ii) image inpainting with a total variation prior and (iii) Poisson image restoration with a synthesis sparsity prior. The associated experimental results will be presented in Section 2.3. In the sequel, note that we put the emphasis on the illustration of the relevance of the AXDA framework to deal with complex models rather than on an optimal image restoration approach dedicated to each problem.

2.2.1 *Unsupervised image deconvolution with a smooth prior*

Problem statement – In this first experiment, we apply the proposed SGS to a classical Bayesian inference problem which involves sampling from a high-dimensional Gaussian posterior distribution with a non-standard covariance matrix (Marnissi et al., 2018). A blurred and noisy image $\mathbf{y} \in \mathbb{R}^d$ (represented as a vector by lexicographic ordering) is observed and assumed to be related to the unknown original image $\boldsymbol{\theta} \in \mathbb{R}^d$ via the Gaussian linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Omega}^{-1}), \tag{2.8}$$

where $\mathbf{H}$ is a $d \times d$ circulant convolution matrix associated to a time/space-invariant blurring kernel. The noise covariance matrix is chosen as diagonal but not necessarily proportional to the identity matrix, i.e., $\boldsymbol{\Omega}^{-1} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. More precisely, we assume that the diagonal elements $\{\sigma_i^2; i \in [d]\}$ of the noise covariance matrix have been randomly drawn according to the mixture $\sigma_i \sim (1-\beta)\delta_{\kappa_1} + \beta\delta_{\kappa_2}$, where $\kappa_1 \ll \kappa_2$. This particular structure for $\boldsymbol{\Omega}^{-1}$ can for instance arise in radar or mobile radio applications (Velayudhan and Paul, 2016; Chang et al., 2016). In those cases, $\boldsymbol{\varepsilon}$ stands for a mixed impulse Gaussian noise and β represents the probability that its variance equals κ_2. The linear inverse problem (2.8) is in generally badly conditioned. To bypass this issue, we adopt the Bayesian framework and consider the smoothing conjugate Gaussian prior distribution (Molina and Ripley, 1989; Molina, Mateos, and Katsaggelos, 2006; Likas and Galatsanos, 2004)

$$\pi(\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0}_d, \left(\gamma \mathbf{L}^T\mathbf{L}\right)^{-1}\right), \tag{2.9}$$

where $\mathbf{L} = \delta\mathbf{I}_d - \tilde{\mathbf{L}}$. The matrix $\tilde{\mathbf{L}}$ is a circulant matrix associated to a Laplacian filter. Since the first eigenvalue of $\tilde{\mathbf{L}}^T\tilde{\mathbf{L}}$ equals zero, we ensure the non-degeneracy and well-posedness of the prior $\pi(\boldsymbol{\theta})$ by introducing the constant diagonal term $\delta\mathbf{I}_d$ with $\delta = 0.1$. We assume here that the hyperparameters κ_1, κ_2, β and γ are unknown. In particular, this implies that the noise standard deviation $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_d]^T$ is also unknown. As such, we want to estimate them jointly with the unknown original image $\boldsymbol{\theta}$. Such an estimation problem is often coined *unsupervised* estimation (Idier, 2008). For these hyperparameters, we adopt the following prior distributions:

$$\pi(\kappa_i^2) = \mathcal{IG}(a, b), i \in [2], \tag{2.10}$$
$$\pi(\beta) = \mathcal{U}([0,1]), \tag{2.11}$$
$$\pi(\gamma) = \mathcal{IG}(a, b). \tag{2.12}$$

Under these prior distributions, the conditional posterior distribution associated to the variable of interest writes

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}, \kappa_1, \kappa_2, \beta, \gamma) = \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{Q}^{-1}\right) \tag{2.13}$$

where

$$\begin{cases} \mathbf{Q} = \mathbf{H}^T\mathbf{\Omega}\mathbf{H} + \gamma\mathbf{L}^T\mathbf{L} & (2.14) \\ \boldsymbol{\mu} = \mathbf{Q}^{-1}\mathbf{H}^T\mathbf{\Omega}\mathbf{y}. & (2.15) \end{cases}$$

Direct sampling according to the posterior distribution (2.13) is a challenging task, mainly due to the presence of the precision matrix $\mathbf{\Omega}$. Indeed, the two terms in (2.14) cannot be diagonalized in the same basis (e.g., Fourier) which leads to computational problems in high dimension. The conditional posterior distributions associated to the unknown hyperparameters present no particular difficulty. As such, we do not write their associated expressions for simplicity and refer the interested reader to the work by Marnissi et al. (2018) for more details.

Proposed approach – To mitigate the sampling issue associated to the posterior of $\boldsymbol{\theta}$, we propose to separate the two terms in the precision matrix (2.14) by relying on the proposed AXDA framework. By setting $\kappa_\rho(\mathbf{z};\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}|\mathbf{H}\boldsymbol{\theta}, \rho^2\mathbf{I}_d)$, the approximate augmented posterior distribution associated to $(\boldsymbol{\theta}, \mathbf{z})$ becomes

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, \boldsymbol{\sigma}, \kappa_1, \kappa_2, \beta, \gamma) \propto \exp\left(-\frac{1}{2}\left[(\mathbf{z}-\mathbf{y})^T\mathbf{\Omega}(\mathbf{z}-\mathbf{y}) + \gamma\left\|\mathbf{L}\boldsymbol{\theta}\right\|^2 + \frac{1}{\rho^2}\left\|\mathbf{z}-\mathbf{H}\boldsymbol{\theta}\right\|^2\right]\right). \quad (2.16)$$

To make the connection with the general expression in (2.2), we have here $p = 2$ and the potentials f_1 and f_2 equal respectively to

$$f_1(\mathbf{A}_1\boldsymbol{\theta};\mathbf{y}) = \frac{1}{2}(\mathbf{A}_1\boldsymbol{\theta}-\mathbf{y})^T\mathbf{\Omega}(\mathbf{A}_1\boldsymbol{\theta}-\mathbf{y}) \quad \text{with } \mathbf{A}_1 = \mathbf{H}, \quad (2.17)$$

$$f_2(\mathbf{A}_2\boldsymbol{\theta}) = \frac{\gamma}{2}\left\|\mathbf{A}_2\boldsymbol{\theta}\right\|^2 \quad \text{with } \mathbf{A}_2 = \mathbf{L}. \quad (2.18)$$

Thanks to the AXDA framework, the two terms in the precision matrix (2.14) are now associated to two different random variables $\boldsymbol{\theta}$ and $\mathbf{z}$. As such, sampling both $\boldsymbol{\theta}$ and $\mathbf{z}$ will be simpler than sampling directly from the original Gaussian posterior (2.13).

Sampling the variable of interest. Under the joint posterior (2.16), the conditional distribution associated to $\boldsymbol{\theta}$ writes

$$\pi_\rho(\boldsymbol{\theta}|\mathbf{z}, \gamma) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}}^{-1}\right) \quad (2.19)$$

where

$$\begin{cases} \mathbf{Q}_\theta = \frac{1}{\rho^2}\mathbf{H}^T\mathbf{H} + \gamma\mathbf{L}^T\mathbf{L} & (2.20) \\ \boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{Q}^{-1}\mathbf{H}^T\frac{\mathbf{z}}{\rho^2}. & (2.21) \end{cases}$$

Since $\mathbf{L}$ and $\mathbf{H}$ are circulant matrices, sampling from this distribution can be efficiently achieved in the Fourier domain. More precisely, $\mathbf{L}$ and $\mathbf{H}$

can be diagonalized in the Fourier domain such that

$$\mathbf{L} = \mathbf{F}^H \mathbf{\Lambda_L} \mathbf{F}, \tag{2.22}$$

$$\mathbf{H} = \mathbf{F}^H \mathbf{\Lambda_H} \mathbf{F}, \tag{2.23}$$

where $\mathbf{F}$ and $\mathbf{F}^H$ are unitary matrices ($\mathbf{F}^H\mathbf{F} = \mathbf{F}\mathbf{F}^H = \mathbf{I}_d$) associated with the Fourier and inverse Fourier transforms. $\mathbf{\Lambda_L}$ and $\mathbf{\Lambda_H}$ are the diagonal counterpart of $\mathbf{L}$ and $\mathbf{H}$ in the Fourier domain, respectively. Using (2.22) and (2.23), the precision matrix $\mathbf{Q}_\theta$ has the form

$$\mathbf{Q}_\theta = \mathbf{F}^H \left(\gamma {\mathbf{\Lambda_L}}^H \mathbf{\Lambda_L} + \frac{1}{\rho^2} {\mathbf{\Lambda_H}}^H \mathbf{\Lambda_H} \right) \mathbf{F}. \tag{2.24}$$

Then, the counterpart of $\mathbf{Q}_\theta$ in the Fourier domain is diagonal and has the form

$$\mathbf{\Lambda}_{\mathbf{Q}_\theta} = \gamma {\mathbf{\Lambda_L}}^H \mathbf{\Lambda_L} + \frac{1}{\rho^2} {\mathbf{\Lambda_H}}^H \mathbf{\Lambda_H}. \tag{2.25}$$

Hence, one can sample from the posterior distribution of $\boldsymbol{\theta}$ by drawing d independent Gaussian samples in the Fourier domain.

Sampling the auxiliary variable. Under the joint (2.16), the posterior associated to $\mathbf{z}$ writes

$$\pi_\rho(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\sigma}, \kappa_1, \kappa_2, \beta) = \mathcal{N}\left(\boldsymbol{\mu}_\mathbf{z}, \mathbf{Q}_\mathbf{z}^{-1}\right) \tag{2.26}$$

where

$$\begin{cases} \mathbf{Q}_\mathbf{z} = \dfrac{1}{\rho^2}\mathbf{I}_d + \mathbf{\Omega} & (2.27) \\ \boldsymbol{\mu}_\mathbf{z} = \mathbf{Q}^{-1}\left(\dfrac{\mathbf{H}\boldsymbol{\theta}}{\rho^2} + \mathbf{\Omega}\mathbf{y}\right). & (2.28) \end{cases}$$

Under this conditional distribution, $\mathbf{z}$ can be efficiently sampled in $\mathbb{R}^d$ since $\mathbf{\Omega}$ was assumed diagonal.

Experimental results associated to this unsupervised image deconvolution problem are presented in Section 2.3.1.

2.2.2 *Image inpainting with a total variation prior*

Problem statement – We illustrate here the benefits 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

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \tag{2.29}$$

where $n < d$, $\mathbf{H} \in \mathbb{R}^{n \times d}$ stands for a decimation binary matrix obtained by taking a subset of rows of the identity matrix $\mathbf{I}_d$ and such that $\mathbf{H}^T\mathbf{H} =$

$\mathbf{I}_n$. 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 (2.29) is ill-posed. To cope with this issue, we assign the isotropic total variation prior distribution to the unknown parameter $\boldsymbol{\theta}$, leading to the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{\|\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\|^2}{2\sigma^2} - \tau \sum_{1\leq i\leq d} \|\mathbf{D}_i\boldsymbol{\theta}\|\right), \tag{2.30}$$

where $\tau > 0$ is a fixed and known regularization parameter and $\mathbf{D}_i\boldsymbol{\theta} \in \mathbb{R}^2$ denotes the two-dimensional discrete gradient applied at pixel i of the image $\boldsymbol{\theta}$, see Chambolle et al. (2010) for more details about the total variation regularization. The presence of the operators $(\mathbf{D}_i; i \in [d])$ and the non-differentiability of the total variation norm rule out the use of common data augmentation schemes and standard simulation-based algorithms (e.g., Hamiltonian and Langevin Monte Carlo methods). Possible surrogates are proximal MCMC methods (Pereyra, 2016; Durmus, Moulines, and Pereyra, 2018) which replace the non-differentiable posterior distribution by a smooth approximation based on the Moreau-Yosida regularization (Moreau, 1965) of the total variation norm. However, the latter does not admit a closed-form expression and iterative routines have been used by Pereyra (2016) and Durmus, Moulines, and Pereyra (2018) to approximate the latter leading to higher computational costs (Chambolle, 2004). We would like to point out that one can avoid these iterative routines by considering separately the ℓ_2 norm and the discrete gradient operator, see for instance (Luu, Fadili, and Chesneau, 2020, Lemma 3.2). In this case, only the ℓ_2 norm needs to be smoothed.

Proposed approach – To mitigate these issues, we propose to rely on the proposed AXDA framework by smoothing the total variation prior with a Gaussian term, leading to the approximate joint posterior density

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:d}|\mathbf{y}) \propto \exp\left(-\frac{\|\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\|^2}{2\sigma^2}\right) \prod_{i=1}^{d} \exp\left(-\tau\|\mathbf{z}_i\| - \frac{\|\mathbf{z}_i - \mathbf{D}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right), \tag{2.31}$$

where for $i \in [d]$, $\mathbf{z}_i \in \mathbb{R}^2$. In the setup considered in Section 2.1, this boils down to setting $b = d+1$, $p = d$, $\mathbf{A}_i = \mathbf{D}_i$ for $i \in [d]$ and $\mathbf{A}_{d+1} = \mathbf{H}$. The associated potential functions write

$$f_i(\mathbf{A}_i\boldsymbol{\theta}) = \tau\|\mathbf{A}_i\boldsymbol{\theta}\| \text{ for } i \in [d], \tag{2.32}$$

$$f_{d+1}(\mathbf{A}_{d+1}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\mathbf{A}_{d+1}\boldsymbol{\theta} - \mathbf{y}\|^2. \tag{2.33}$$

Let decompose the two-dimensional discrete gradient at pixel i as $\mathbf{D}_i = [(\mathbf{D}^{(1)}\boldsymbol{\theta})_i; (\mathbf{D}^{(2)}\boldsymbol{\theta})_i]$ where $\mathbf{D}^{(1)}$, $\mathbf{D}^{(2)}$ represent the two first-order forward finite difference operators in horizontal and vertical directions, respectively. Then, by defining $\mathbf{D} = [\mathbf{D}^{(1)}; \mathbf{D}^{(2)}]$ and since $\ker(\mathbf{H}) \cap \ker(\mathbf{D}) = \{\mathbf{0}_d\}$, the posterior density $\pi_\rho(\boldsymbol{\theta})$ under (2.31) is proper. By relying on this joint approximate posterior density, the inference is now simplified and can be conducted easily with Algorithm 1. In the following

The notation $\ker(\mathbf{D})$ refers to the null space of the matrix $\mathbf{D}$. For this matrix, $\ker(\mathbf{D}) = \{\mathbf{0}_d\} \cup \{\lambda\mathbf{1}_d \in \mathbb{R}^d; \lambda \in \mathbb{R}\}$. Since the matrix $\mathbf{H}$ only contains zeros and ones, constant vectors of the form $\lambda\mathbf{1}_d$ cannot belong to $\ker(\mathbf{H})$. This leads to $\ker(\mathbf{H}) \cap \ker(\mathbf{D}) = \{\mathbf{0}_d\}$.

paragraphs, we detail the conditional sampling steps of SGS.

Sampling the auxiliary variables. For $i \in [d]$, the conditional distribution associated to the auxiliary variable $\mathbf{z}_i$ admits the density

$$\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}) \propto \exp\left(-\tau \|\mathbf{z}_i\| - \frac{\|\mathbf{z}_i - \mathbf{D}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right). \tag{2.34}$$

This conditional distribution can be sampled exactly by using data augmentation. Indeed, one can re-write the distribution involving the non-differentiable potential $\|\cdot\|$ as a mixture of normal and gamma distributions, see Kyung et al. (2010, Section 3.1). Hence, sampling from (2.34) can be performed with the following two steps

$$\text{Draw } \frac{1}{\gamma_i} \mid \mathbf{z}_i \sim \begin{cases} \text{InverseGaussian}\left(\dfrac{\tau}{\|\mathbf{z}_i\|}, \tau^2\right) & \text{if } \|\mathbf{z}_i\| > 0 \\ \text{InverseGaussian}\left(\dfrac{3}{2}, \dfrac{\tau^2}{2}\right) & \text{if } \|\mathbf{z}_i\| = 0 \end{cases} \tag{2.35}$$

$$\text{Draw } \mathbf{z}_i \mid \gamma_i, \boldsymbol{\theta} \sim \mathcal{N}\left(\frac{\gamma_i \mathbf{D}_i\boldsymbol{\theta}}{\rho^2+\gamma_i}, \frac{\rho^2\gamma_i}{\rho^2+\gamma_i}\mathbf{I}_2\right). \tag{2.36}$$

Note that these last d sampling steps (associated to the $\mathbf{z}_i$'s) can be performed efficiently by "vectorizing" them.

Sampling the parameter of interest. On the other hand, the conditional distribution associated to the image to recover $\boldsymbol{\theta}$ admits the density

$$\pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:d}, \mathbf{y}) \propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{H}\boldsymbol{\theta}\|^2}{2\sigma^2} - \sum_{i=1}^{d} \frac{\|\mathbf{z}_i - \mathbf{D}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right). \tag{2.37}$$

The distribution (2.37) is a non-degenerate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ where

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left(\rho^{-2}\mathbf{D}^T\mathbf{D} + \sigma^{-2}\mathbf{H}^T\mathbf{H}\right)^{-1} \tag{2.38}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\left(\sigma^{-2}\mathbf{H}^T\mathbf{y} + \rho^{-2}\sum_{i=1}^{d}\mathbf{D}_i^T\mathbf{z}_i\right). \tag{2.39}$$

Under periodic boundary conditions for $\boldsymbol{\theta}$, the matrix $\mathbf{D}^T\mathbf{D}$ is block circulant and hence diagonalizable in the Fourier domain, see Wang et al. (2008) for more details. 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 and sample from this high-dimensional Gaussian distribution. Let $\eta \left\|\mathbf{H}^T\mathbf{H}\right\| < \sigma^2$ where $\|\mathbf{M}\|$ is the spectral norm of the matrix

$\mathbf{M}$. Then, we have the following two-step sampling scheme

$$\text{Draw } \mathbf{v} \mid \boldsymbol{\theta} \sim \mathcal{N}\left(\left(\frac{\mathbf{I}_d}{\eta} - \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\right)\boldsymbol{\theta}, \frac{\mathbf{I}_d}{\eta} - \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\right), \tag{2.40}$$

$$\text{Draw } \boldsymbol{\theta} \mid \mathbf{v}, \mathbf{y}, \mathbf{z}_{1:d} \sim \mathcal{N}\left(\boldsymbol{\mu}'_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}'_{\boldsymbol{\theta}}\right), \tag{2.41}$$

where

$$\boldsymbol{\Sigma}'_{\boldsymbol{\theta}} = \left(\frac{\mathbf{I}_d}{\eta} + \frac{\mathbf{D}^T\mathbf{D}}{\rho^2}\right)^{-1}, \tag{2.42}$$

$$\boldsymbol{\mu}'_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}'_{\boldsymbol{\theta}}\left(\mathbf{v} + \frac{\mathbf{H}^T}{\sigma^2}\mathbf{y} + \frac{1}{\rho^2}\sum_{i=1}^{d}\mathbf{D}_i^T\mathbf{z}_i\right). \tag{2.43}$$

Experimental results associated to this image inpainting problem are presented in Section 2.3.2.

2.2.3 *Poisson image restoration with a frame-based synthesis approach*

For the problems considered in Sections 2.2.1 and 2.2.2, there already exist state-of-the-art MCMC approaches to tackle such problems (Marnissi et al., 2018; Durmus, Moulines, and Pereyra, 2018). Still, the numerical results in Sections 2.3.1 and 2.3.2 will show that SGS competes with and, in some cases, even improves upon those existing methods.

Problem statement – We consider the observation of some image $\mathbf{y} \in \mathbb{N}^n$, blurred and contaminated by Poisson noise. We assume that each individual observation $(y_i; i \in [n])$ corresponds to an independent realization of a Poisson random variable, that is for all $i \in [n]$,

$$y_i|\mathbf{x} \overset{\text{i.i.d.}}{\sim} \mathcal{P}\left([\mathbf{H}\mathbf{x}]_i\right), \tag{2.44}$$

where $\mathbf{x} \in \mathbb{R}_+^m$ stands for an unknown image to recover and $\mathbf{H} \in \mathbb{R}^{n\times m}$ is an operator related to the point spread function. We consider here a frame-based synthesis approach and assume that the image $\mathbf{x}$ can be written according to the linear generative model

$$\mathbf{x} = \boldsymbol{\Phi}\boldsymbol{\theta} = \sum_{i=1}^{d}\theta_i\boldsymbol{\phi}_i, \tag{2.45}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d] \in \mathbb{R}^{m\times d}$ is a so-called *dictionary* whose columns are called *atoms*, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is the vector gathering the representation coefficients of $\mathbf{x}$ in the dictionary $\boldsymbol{\Phi}$ (Starck, Murtagh, and Fadili, 2015). In the sequel, we will build upon the synthesis representation (2.45) of the original image $\mathbf{x}$ to exploit the sparsity of this image in $\boldsymbol{\Phi}$, that is, we expect to represent $\mathbf{x}$ by only using a superposition of few atoms.

We also assume that for all $\mathbf{x} \in \mathbb{R}_+^m$, $\mathbf{H}\mathbf{x} \in \mathbb{R}_+^n$. Note that we do not consider any background noise here for simplicity. Nevertheless, the proposed methodology can be easily generalized to this scenario. In order to

make an easier link with the proposed approach, the likelihood distribution is re-written as follows

$$\pi(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left(-f_1(\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta};\mathbf{y})\right), \tag{2.46}$$

where $f_1(\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta};\mathbf{y}) = \sum_{i=1}^{n} \xi([\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i; y_i)$ with

$$\xi([\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i; y_i) = \begin{cases} -y_i \log\left([\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i\right) + [\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i & \text{if } [\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i > 0, \\ 0 & \text{if } [\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i = 0, \\ \infty & \text{otherwise.} \end{cases} \tag{2.47}$$

In the following, we will use the AXDA framework and introduce some auxiliary variables $(z_i; i \in [n])$ standing for replicates of the variables $([\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}]_i; i \in [n])$. To guarantee the positivity of these auxiliary variables, we modify (2.47) such that (Figueiredo and Bioucas-Dias, 2010)

$$f_1(\mathbf{z};\mathbf{y}) = \sum_{i=1}^{n} -y_i \log(z_i) + z_i + \iota_{\mathbb{R}_+}(z_i). \tag{2.48}$$

Following a Bayesian approach (Robert, 2001), the prior distribution is defined as follows for all $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-f_2(\boldsymbol{\theta}) - f_3(\boldsymbol{\Phi}\boldsymbol{\theta})\right), \tag{2.49}$$

with

$$f_2(\boldsymbol{\theta}) = \tau \left\|\boldsymbol{\theta}\right\|_1 \text{ and } f_3(\boldsymbol{\Phi}\boldsymbol{\theta}) = \iota_{\mathbb{R}_+^m}(\boldsymbol{\Phi}\boldsymbol{\theta}), \tag{2.50}$$

where τ is a positive parameter. The functions f_2 and f_3 being convex, the potential $f_2 + f_3$ is also convex and the prior $\pi(\boldsymbol{\theta})$ is log-concave. The function f_2 encodes our prior knowledge that $\mathbf{x}$ is sparse in $\boldsymbol{\Phi}$ while the function f_3 guarantees the non-negativity of the original image $\mathbf{x}$ since the latter can be viewed as an intensity. By the application of Bayes' rule, the posterior distribution writes

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-f_1(\mathbf{H}\boldsymbol{\Phi}\boldsymbol{\theta}) - f_2(\boldsymbol{\theta}) - f_3(\boldsymbol{\Phi}\boldsymbol{\theta})\right), \tag{2.51}$$

and has the following properties stated by Proposition 9.

Proposition 9. *1. The posterior distribution (2.51) is log-concave.*

2. *The potential function $f = f_1 + f_2 + f_3$ associated to π is proper, lower semi-continuous, coercive and convex. Additionally, if $y_i \neq 0$ for all $i \in [n]$, f is strictly convex.*

3. *The negative log-likelihood function f_1 is not gradient-Lipschitz continuous.*

Proof. The proof of these properties follows directly from the work by Figueiredo and Bioucas-Dias (2010). □

Up to our knowledge, no previous work considered the direct or approximate sampling from the posterior distribution defined in (2.51). Following Proposition 9, $-\log \pi$ is convex and possibly not differentiable. In this setting, since $-\log \pi$ is not differentiable, one could not resort to Hamiltonian Monte Carlo methods (Duane et al., 1987) to sample from (2.51). These methods were for instance used with $f_2 = f_3 = 0$ in the work by Pedemonte, Catana, and Van Leemput (2015). Then, one could think of using proximal MCMC approaches (Pereyra, 2016; Durmus, Moulines, and Pereyra, 2018) to tackle the non-differentiability of the potential function associated to π thanks to Moreau-Yosida regularization (Moreau, 1965). However, these approaches require the existence of a smooth gradient-Lipschitz continuous term in the potential function which is not the case in the problem considered, see Property 3 in Proposition 9. Note that it is possible to tackle this issue by using the Anscombe variance stabilizing transform (Anscombe, 1948) but the performance of the latter is known to degrade in low intensity regimes (Dupé, Fadili, and Starck, 2012). Additionally, proximal MCMC algorithms assume that the proximity operator associated to the non-smooth potential, here $f_2 + f_3$, is available. This is not the case for the considered potentials (Dupé, Fadili, and Starck, 2009). Therefore, one has to resort to a more complicated scheme such as the splitting method of maximal monotone operators to compute this proximity operator (Eckstein and Bertsekas, 1992; Lions and Mercier, 1979).

Proposed approach – We tackle these issues by relying on the AXDA framework and SGS. We consider a full-splitting strategy by introducing three auxiliary variables associated to each potential $(f_i; i \in [3])$ of the posterior (2.51) such that

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{y}) \propto \prod_{i=1}^{3} \exp\left(-f_i(\mathbf{z}_i) - \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2}{2\rho^2}\right), \qquad (2.52)$$

where $\mathbf{A}_1 = \mathbf{H}\boldsymbol{\Phi}$, $\mathbf{A}_2 = \mathbf{I}_d$ and $\mathbf{A}_3 = \boldsymbol{\Phi}$.

Sampling the auxiliary variables. In this scenario, the full conditional distribution associated to each auxiliary variable $(\mathbf{z}_i; i \in [3])$ writes

$$\pi_\rho(\mathbf{z}_i, |\mathbf{y}, \boldsymbol{\theta}) \propto \exp\left(-f_i(\mathbf{z}_i) - \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2}{2\rho^2}\right). \qquad (2.53)$$

The latter is log-concave with a smooth gradient Lipschitz term (related to the Gaussian term) and a potentially non-smooth term f_i. Concerning $\mathbf{z}_1$ and $\mathbf{z}_3$, the proximity operators of f_1 and f_3 are available in closed-form and hence proximal MCMC methods can be resorted to sample from this conditional (Figueiredo and Bioucas-Dias, 2010; Dupé, Fadili, and Starck, 2009). On the other hand, the conditional distribution of $\mathbf{z}_2$ can be sampled exactly by exploiting the mixture representation of the Laplacian distribution (Park and Casella, 2008).

Sampling the variable of interest. The full conditional distribution associ-

ated to $\boldsymbol{\theta}$ is Gaussian and writes

$$\pi_\rho(\boldsymbol{\theta}, |\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) \propto \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}}^{-1}), \tag{2.54}$$

with

$$\begin{cases} \mathbf{Q}_{\boldsymbol{\theta}} &= \rho^{-2}(\boldsymbol{\Phi}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\Phi} + \mathbf{I}_d + \boldsymbol{\Phi}^T\boldsymbol{\Phi}) \\ \boldsymbol{\mu}_{\boldsymbol{\theta}} &= (\boldsymbol{\Phi}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\Phi} + \mathbf{I}_d + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^T\mathbf{H}^T\mathbf{z}_1 + \mathbf{z}_2 + \boldsymbol{\Phi}^T\mathbf{z}_3). \end{cases} \tag{2.55}$$

This Gaussian distribution can be easily sampled if $\mathbf{H}$ is associated to a periodic convolution. Indeed, in this case $\mathbf{H}$ is block circulant and hence diagonalizable in the Fourier domain. Using the Sherman-Morrison-Woodbury matrix inversion formula, it follows that

$$(\boldsymbol{\Phi}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\Phi} + \mathbf{I}_d + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} = (\boldsymbol{\Phi}^T(\mathbf{H}^T\mathbf{H} + \mathbf{I}_d)\boldsymbol{\Phi} + \mathbf{I}_d)^{-1} \tag{2.56}$$

$$= \mathbf{I}_d - \boldsymbol{\Phi}^T\left(\mathbf{I}_d + (\mathbf{H}^T\mathbf{H} + \mathbf{I}_d)^{-1}\right)^{-1}\boldsymbol{\Phi} \tag{2.57}$$

$$= \mathbf{I}_d - \boldsymbol{\Phi}^T\mathbf{F}^H\left(\mathbf{I}_d + (|\mathbf{D}|^2 + \mathbf{I}_d)^{-1}\right)^{-1}\mathbf{F}\boldsymbol{\Phi}, \tag{2.58}$$

where $\mathbf{H} = \mathbf{F}^H\mathbf{D}\mathbf{F}$ with $\mathbf{F}$ and $\mathbf{F}^H$ being unitary matrices associated with the Fourier and inverse Fourier transforms. This implies that sampling from (2.54) can be performed in $\mathcal{O}(d\log(d))$ operations using the fast Fourier transform implementation of $\mathbf{F}$ and $\mathbf{F}^H$.

Experimental results associated to this image inpainting problem are presented in Section 2.3.3.

2.3 Experiments

This section aims at illustrating experimentally the main benefits of SGS on the canonical Bayesian inference problems introduced in Section 2.2. We do not pretend to find the best solutions to these problems but rather to point out that the proposed approach can be used to tackle them efficiently. These problems are typically solved via MAP estimation with optimization-based approaches since the latter deliver accurate solutions very quickly. We show here that SGS stands for an efficient simulation-based approach, that could be subsequently resorted to conduct challenging statistical analyses (e.g., performing model choice, deriving credibility intervals or computing Bayesian estimators) that are beyond the scope of classical optimization approaches. All the experiments have been carried out on a Dell Latitude 7390 laptop equipped with an Intel(R) Core(TM) i5-8250U 1.60 GHz processor, with 16.0 GB of RAM, running Windows 10. The code associated to these experiments is available online at https://github.com/mvono.

2.3.1 *Unsupervised image deconvolution with a smooth prior*

Experimental setting – The unsupervised image deconvolution problem introduced in Section 2.2.1 and also addressed by Marnissi et al. (2018) is considered. The parameters of the inverse Gamma prior distributions $\pi(\kappa_1^2)$, $\pi(\kappa_2^2)$ and $\pi(\gamma)$ are set to $(a, b) = (0.1, 0.1)$ such that these prior become diffuse and encode no prior knowledge. Figure 2.1 presents the 512×512 ($d = 262,144$) original gray-level BOAT image used for this experiment and its associated blurred and noisy observation. The proposed SGS is compared to the reversible jump perturbation-optimization (RJPO) algorithm (Gilavert, Moussaoui, and Idier, 2015) and to the algorithms coined AuxV1 and AuxV2 proposed by Marnissi et al. (2018). The tolerance parameter associated to SGS has been set to $\rho = 20$. RJPO has been run using conjugate gradient (CG) algorithm as the required linear solver whose tolerance has been adapted to reach an acceptance rate of 0.9. The number of burn-in iterations has been set to $T_{\mathrm{bi}} = 200$ for AuxV1, RJPO and SGS and to $T_{\mathrm{bi}} = 2200$ for AuxV2 (due to its slower mixing properties, see below). For each MCMC algorithm, 800 samples obtained after the burn-in period have been used. The number of iterations T_{MC} and T_{bi} were empirically chosen by graphically inspecting the behavior of the Markov chains produced by the samplers.

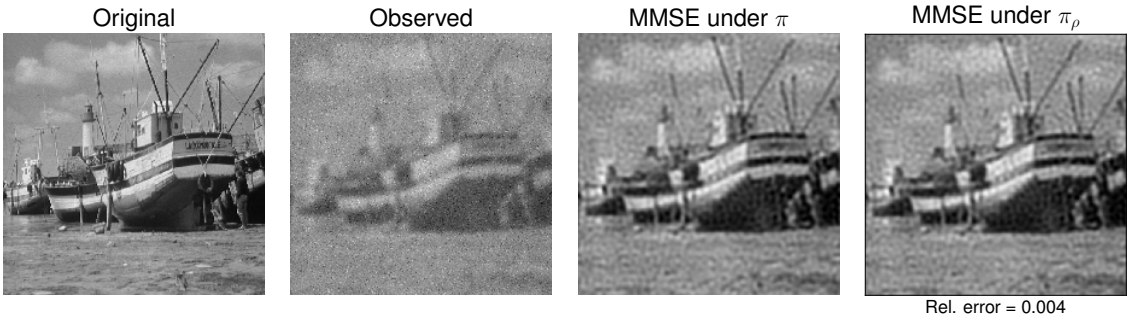


Figure 2.1: From left to right: (i) original image, (ii) blurred and noisy observation, (iii) MMSE estimate of $\boldsymbol{\theta}$ under π, i.e., $\mathbb{E}_{\pi}(\boldsymbol{\theta})$ and (iv) MMSE estimate of $\boldsymbol{\theta}$ under π, i.e., $\mathbb{E}_{\pi_\rho}(\boldsymbol{\theta})$. The relative error stands for the relative error between the approximate and true MMSE estimates.

Deconvolution results – Figure 2.1 shows the minimum mean-square error (MMSE) estimator computed with SGS along with the true MMSE under π. One can denote that the MMSE under π_ρ is in agreement with the image deconvolution task and is very close to the true MMSE (relative error of 0.4%). The dot/circle patterns we can observe in the reconstructed images correspond to oscillation artifacts due to the Tikhonov prior distribution. Table 2.1 measures quantitatively the restoration performances of the different approaches by assessing their signal-to-noise ratio (SNR) and their peak signal-to-noise ratio (PSNR) given by

$$\mathrm{SNR} = 10 \log_{10} \frac{\|\boldsymbol{\theta}\|^2}{\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right\|^2} \tag{2.59}$$

$$\mathrm{PSNR} = 10 \log_{10} \frac{255^2}{d^{-1}\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right\|^2}, \tag{2.60}$$

The MMSE is given by the posterior mean and is defined by

$$\mathbb{E}_{\pi}(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}.$$

where $\hat{\boldsymbol{\theta}}$ denotes the MMSE estimate of $\boldsymbol{\theta}$ approximated by empirical averages of the samples generated by the MCMC algorithms. The standard deviation associated to these results is the same for the different methods and is equal to 0.02. One can denote that all the algorithms share similar performance. However, we emphasize that the computational efficiency of each algorithm can differ widely as shown in the following paragraph.

method	SNR (dB)	PSNR (dB)
SGS	17.57	22.92
AuxV1	17.57	22.92
AuxV2	17.58	22.93
RJPO	17.57	22.91

Table 2.1: Performance results associated to the MMSE.

Computational efficiency – We assess here the efficiency of each approach. After the burn-in period, we measure this efficiency by building upon the effective sample size ratio per second (ESSR) defined as

$$\mathrm{ESSR}(\vartheta) = \frac{1}{T_1}\frac{\mathrm{ESS}(\vartheta)}{T} = \frac{1}{T_1\left(1+2\sum_{t=1}^{\infty}\rho_t(\vartheta)\right)} \tag{2.61}$$

where T_1 is the CPU time in seconds required to draw one sample, T is the number of available samples after the burn-in period and $\rho_t(\vartheta)$ is the lag-t autocorrelation of a scalar parameter ϑ. For an MCMC sampler, the ESSR gives an estimate of the equivalent number of i.i.d. samples that can be drawn in one second (Kass et al., 1998; Liu, 2001). A variant of the ESSR has for instance been used by Gilavert, Moussaoui, and Idier (2015) in order to measure the efficiency of an MCMC algorithm dedicated to Gaussian sampling. In this experiment, we choose to measure the ESSR by using the Markov chain associated to the scalar hyperparameter γ. As in the statistical software STAN (Carpenter et al., 2017), we truncated the infinite sum in (2.61) at the first negative ρ_t.

Table 2.2 compares the ESSR for the four MCMC approaches after the burn-in period. One can denote that AuxV1 admits the best mixing properties which comes at no surprise since this approach is specifically dedicated to Gaussian sampling. On the other hand, SGS compares favorably with AuxV1 while being far more general. SGS is also more efficient than AuxV2 and RJPO although the latter methods are dedicated to Gaussian sampling. The cost per iteration of RJPO is very high due to the number of conjugate gradient iterations (155 on average after the burn-in period) performed at each iteration. Note that RJPO could be accelerated with a preconditioned conjugate gradient by using circulant preconditioners, for instance.

Hyperparameters estimation – Figure 2.2 shows the trace plots associated to the four unknown hyperparameters. One can denote that the four methods indeed manage to converge towards the true values of the hyperparameters $\beta = 0.35$, $\kappa_1 = 13$ and $\kappa_2 = 40$. Concerning the regularization

method	ESSR	T_1 [seconds]
SGS	7.5	0.13
AuxV1	8.1	0.12
AuxV2	6.8	0.15
RJPO	0.20	5.11

Table 2.2: Efficiency of the MCMC samplers. The ESSR has been computed using the Markov chain associated to γ as scalar summary. T_1 stands for CPU time in seconds required to draw one sample.

hyperparameter γ which is not related to the synthetic generation of the observed image, one can note that its value in the stationary regime differs between SGS and exact samplers. Table 2.3 provides point-wise MMSE estimates and associated standard deviations.

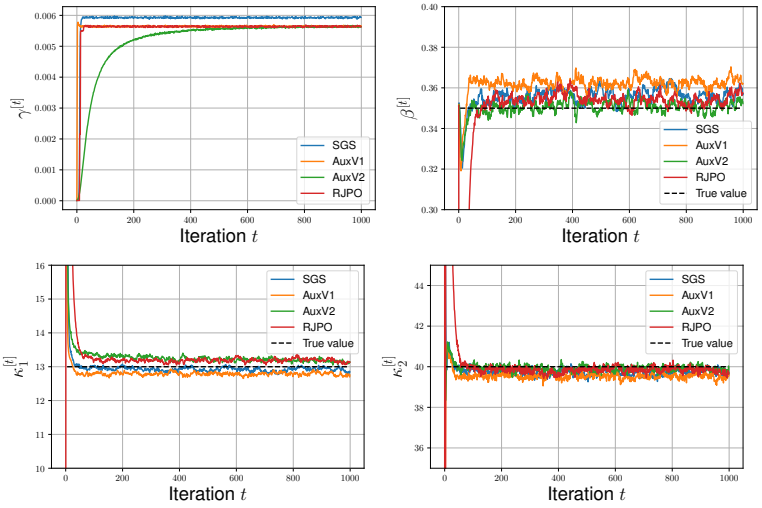


Figure 2.2: Trace plots associated to the unknown hyperparameters. From top left to bottom right: (i) γ, (ii) β, (iii) κ_1 and (iv) κ_2. The true value associated to the last three hyperparameters is shown via a black dashed line.

method	γ	β	κ_1	κ_2
True value	-	0.35	13	40
SGS	5.9 $\times 10^{-3}$	0.356 ($\pm$ 0.003)	12.92 ($\pm$ 0.06)	39.71 ($\pm$ 0.15)
AuxV1	5.6 $\times 10^{-3}$	0.362 ($\pm$ 0.003)	12.79 ($\pm$ 0.05)	39.53 ($\pm$ 0.14)
AuxV2	5.6 $\times 10^{-3}$	0.355 ($\pm$ 0.003)	13.19 ($\pm$ 0.05)	39.84 ($\pm$ 0.14)
RJPO	5.7 $\times 10^{-3}$	0.352 ($\pm$ 0.003)	13.05 ($\pm$ 0.09)	39.87 ($\pm$ 0.14)

Table 2.3: MMSE estimates ($\pm$ standard deviation) of the four unknown hyperparameters.

Overall, this first experiment shows that the proposed SGS can compete with efficient algorithms designed only for this type of sampling problems (e.g., AuxV1). Additionally, it proves to be more efficient than algorithms designed for wider Gaussian sampling tasks (e.g., AuxV2 and RJPO). The performance of the proposed approach is strengthened by the fact that SGS has also demonstrated to be more efficient than state-of-the-art MCMC algorithms designed to sample from other types of distributions, such as log-concave densities, as illustrated in the next sections.

2.3.2 *Image inpainting with a total variation prior*

Experimental setting – The image inpainting problem introduced in Section 2.2.2 and also addressed by Afonso, Bioucas-Dias, and Figueiredo (2010) with optimization methods is considered here. Figure 2.3 presents

the 256×256 ($d = 65,536$) original gray-level CAMERAMAN image used for this experiment. The observation vector denoted $\mathbf{y}$ consists of 40% randomly selected pixels of the original image $\boldsymbol{\theta}$, corrupted by a white Gaussian noise with a noise variance $\sigma^2 = 0.39$ which leads to a SNR of 40 dB. Figure 2.3 (top right) shows the observed image where the missing pixels have been depicted in white. The fixed regularization parameter τ is set manually to $\tau = 0.2$.

The proposed SGS algorithm is compared with the proximal Moreau-Yosida unadjusted Langevin algorithm (MYULA), specifically designed to sample approximately from possibly non-smooth log-concave distributions (Durmus, Moulines, and Pereyra, 2018). The tolerance parameters associated to these two approximate samplers have been manually set to $\rho = \sigma$ (for SGS) and $\lambda = 4\gamma = \sigma^2$ (for MYULA) following the guidelines of Durmus, Moulines, and Pereyra (2018). As highlighted in Section 2.2.2, the iterative Chambolle's algorithm (with 20 iterations here) is used within MYULA to compute the proximity operator of the total variation norm (Chambolle, 2004). In order to assess the bias induced by SGS, we also run the Metropolis-adjusted version of MYULA, called MYMALA, which admits the initial target distribution π as invariant distribution.

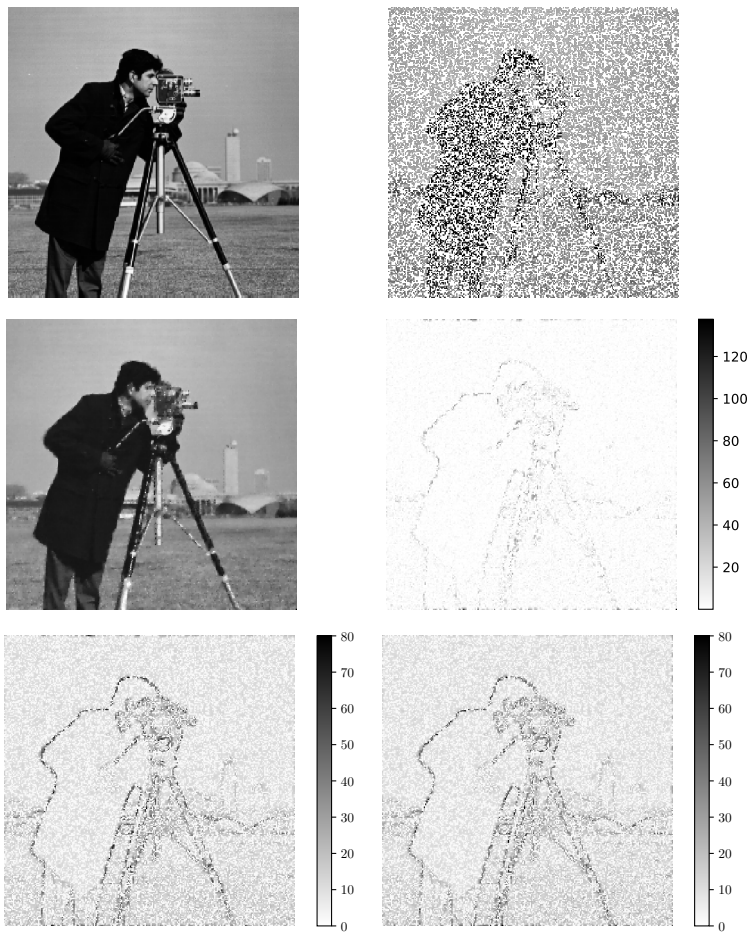


Figure 2.3: From top left to bottom right: (i) original image, (ii) damaged and noisy observation, (iii) MMSE estimate of $\boldsymbol{\theta}$ under π_ρ, i.e., $\mathbb{E}_{\pi_\rho}(\boldsymbol{\theta})$, (iv) absolute pointwise difference $|\mathbb{E}_{\pi}(\boldsymbol{\theta}) - \mathbb{E}_{\pi_\rho}(\boldsymbol{\theta})|$, (v) 90% credibility intervals obtained with SGS and (vi) 90% credibility intervals obtained with MYULA.

Restoration results – For each algorithm, $20,000$ samples obtained after an appropriate burn-in period (which depends on the mixing properties of

each MCMC sampler) are used to assess their performance and possible bias. In the following, we use these samples to compute MMSE estimators, 90% credibility intervals and scalar summaries associated to highest posterior density (HPD) regions.

Figure 2.3 shows the MMSE estimator computed with SGS (middle left) and its absolute pointwise bias w.r.t. the true MMSE estimator under π (middle right). One can denote that the MMSE estimate under π_ρ is indeed in agreement with the image reconstruction task, visually close to the true MMSE estimate and that its main differences with the latter are located near the boundaries of objects in the image where there is more uncertainty, see the bottom images of Figure 2.3 which represent the 90% credibility intervals obtained with SGS and MYULA. Table 2.4 complements these visual inference results by reporting and comparing quantitative measures related to the MMSE obtained with SGS. Since the ground truth $\boldsymbol{\theta}_{\text{true}}$ is known, the performance of this estimator is measured by computing the improvement in signal-to-noise ratio (ISNR) and the mean-square error (MSE) defined as

$$\text{ISNR} = 10\log_{10}\left(\frac{\|\boldsymbol{\theta}_{\text{true}} - \mathbf{y}\|^2}{\|\boldsymbol{\theta}_{\text{true}} - \mathbb{E}_{\pi_\rho}(\boldsymbol{\theta})\|^2}\right) \tag{2.62}$$

$$\text{MSE} = \frac{1}{d}\|\boldsymbol{\theta}_{\text{true}} - \mathbb{E}_{\pi_\rho}(\boldsymbol{\theta})\|^2. \tag{2.63}$$

One can denote that the MMSE estimators obtained with SGS and MYULA are comparable in terms of ISNR and MSE. More importantly, they have a similar relative error w.r.t. the true MMSE estimator (under π) of the order of 4% showing that SGS is able to provide accurate and meaningful results in terms of posterior mean inference. To emphasize the correctness of the proposed approach beyond the comparison between pointwise estimates, we also pay attention to the comparison between posterior credibility sets induced by both π and π_ρ. To this purpose, we consider the highest posterior density region given by

method	ISNR (dB)	MSE	relative error w.r.t. $\mathbb{E}_\pi(\boldsymbol{\theta})$
SGS	18.13	166	0.04
MYULA	18.23	162	0.04

Table 2.4: Performance results associated to the MMSE estimator. The last column reports the relative error between the MMSE estimator computed with each method and the MMSE estimator under the initial target distribution π.

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

where $\gamma_\alpha \in \mathbb{R}$ is such that $\int_{\mathcal{C}_\alpha^\star} \pi(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} = 1-\alpha$ and $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{H}\boldsymbol{\theta}\|^2/(2\sigma^2) + \tau\sum_{1\leq i\leq d}\|\mathbf{D}_i\boldsymbol{\theta}\|$ is the potential function associated to $\pi(\boldsymbol{\theta}|\mathbf{y})$.

Figure 2.4 shows the different values of the scalar summary estimated using MYMALA, SGS and MYULA for $\alpha \in [0.01, 0.99]$. It is denoted by γ_α for the exact sampler MYMALA and by $\tilde{\gamma}_\alpha$ for approximate samplers (SGS and MYULA). Note that the approximation error associated to γ_α lies between 0.1% and 0.3% depending on the value of α, which supports the use of SGS to conduct Bayesian uncertainty analysis in this problem.

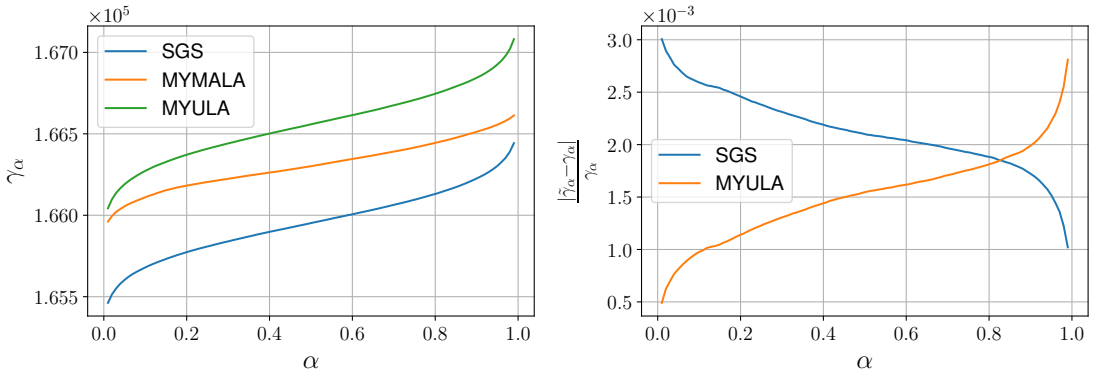


Figure 2.4: (left) Threshold values γ_α associated to SGS, MYULA and MYMALA; (right) relative error between the threshold value estimated with MYMALA denoted γ_α and the one estimated with approximate approaches (SGS and MYULA) denoted $\tilde{\gamma}_\alpha$.

Efficiency of SGS – We now assess and compare the efficiency of SGS by analyzing the convergence and mixing properties of this MCMC sampler. After the burn-in period, we measure this efficiency by building upon the ESSR. Table 2.5 reports the ESSR associated to SGS and MYULA. One can denote that SGS has better mixing properties than MYULA in its stationary regime which confirms the interest of AXDA and the proposed Gibbs sampling procedure for such image processing tasks. To complement this analysis in the stationary regime, we also compute trace plots illustrating the behavior of SGS and MYULA in their transient regime. To this purpose, we use the scalar summary $f(\boldsymbol{\theta})$ which has been shown to concentrate sharply on the typical set $f(\boldsymbol{\theta}) \approx \mathbb{E}(f(\boldsymbol{\theta}))$ (Pereyra, 2019). Figure 2.5 shows the convergence of the Markov chains associated to SGS and MYULA towards this typical set. One can see that SGS explores the typical set roughly 3 times faster than MYULA in terms of number of iterations while having a lower computational cost per iteration. This demonstrates again that SGS stands for an efficient sampling alternative compared to the state-of-the-art MYULA.

method	ESSR	T_1 [seconds]
SGS	0.22	3.83
MYULA	0.15	4.22

Table 2.5: Relative efficiency of SGS compared to MYULA. The ESSR has been computed using $U(\boldsymbol{\theta})$ as scalar summary. T_1 stands for CPU time in seconds required to draw one sample.

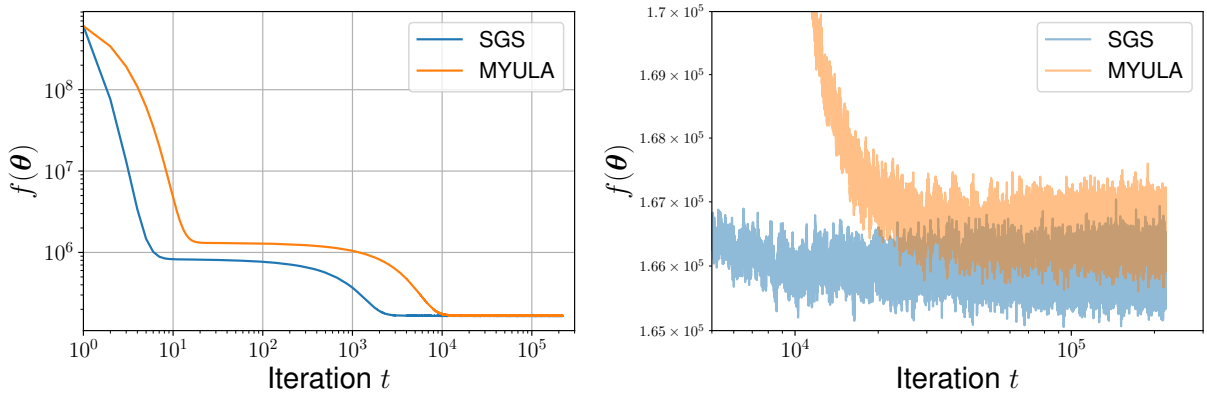


Figure 2.5: Convergence to the typical set of the posterior distribution π for SGS and MYULA.

2.3.3 *Poisson image restoration with a frame-based synthesis approach*

Experimental setting – The Poisson image restoration problem introduced in Section 2.2.3 is considered here. Figure 2.3 depicts the 128×128 ($d = 16,384$) original gray-level NEURON image which will be used for this experiment. In the following, the maximum intensity of this image denoted M will be scaled according to two different values ($M = 30$ and $M = 100$) in order to assess the efficiency of the proposed approach in different scenarios. This image is artificially damaged via the matrix $\mathbf{H}$ related to a Gaussian blurring kernel and corrupted by a Poisson noise. This yields the observation $\mathbf{y}$ shown in Figure 2.3. Reconstructing this image by taking into account the Poisson likelihood is particularly important when the maximum intensity M is low. In this case, this likelihood cannot be approximated by a Gaussian likelihood.

In the sequel, the dictionary $\mathbf{\Phi}$ will be chosen as the Haar wavelet frame with four levels. This poor man's wavelet has been chosen to illustrate the use of a frame-based approach although more sophisticated frames can be considered. Since piecewise constant images (e.g., the NEURON image) are sparse in the Haar wavelet domain, using this representation is similar to the TV regularization (Steidl et al., 2004). Both the regularization parameter τ and the tolerance parameter of SGS ρ have been set according to the maximum intensity level M. Although it cannot be directly applied to sample from the target posterior distribution (2.51) (see Section 2.2.3), the proximal MCMC algorithm MYULA has also been implemented by using the Anscombe variance stabilizing transform (VST) and Douglas-Rachford splitting method to compute the proximity operator of $f_2 + f_3$. Nethertheless, note that a generalized version of MYULA that can be directly applied to this image restoration problem has been recently proposed by Luu, Fadili, and Chesneau (2020). Future works will be dedicated to compare SGS and this sampling algorithm.

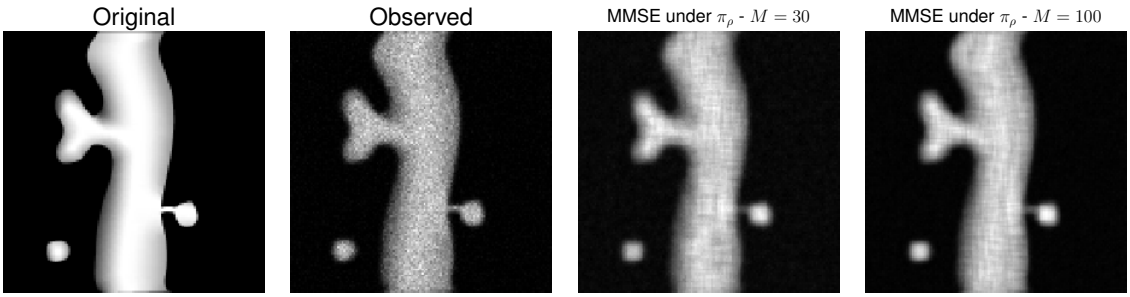


Figure 2.6: From left to right: (i) original image, (ii) blurred and noisy observation, (iii) MMSE estimate of $\boldsymbol{\theta}$ under π_ρ with $M = 30$ and (iv) MMSE estimate of $\boldsymbol{\theta}$ under π_ρ with $M = 100$.

Restoration results – We run each MCMC algorithm with $T = 10^5$ iterations from which 10^4 samples are used to approximate MMSE estimators. Since the ground-truth is known, the performance of each method has been assessed using the mean absolute error (MAE $= d^{-1} \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|_1$) and its normalized version (norm. MAE) w.r.t. the intensity level M. Note that the MAE is particularly relevant for Poissonian image restoration since it is related to other distances (Dupé, Fadili, and Starck, 2009; Barron and Cover, 1991). Figure 2.6 shows MMSE estimators computed using the

samples generated by SGS for $M = 30$ and $M = 100$. One can denote that these point-wise estimators are indeed coherent with the reconstruction task. As emphasized before, although these restoration results could have been improved by considering more sophisticated frames $\boldsymbol{\Phi}$, the goal of this experiment is to show that the proposed approach can be applied to Poisson image restoration without resorting to multiple approximations as in MYULA. Table 2.6 complements these visual results by showing the performances of SGS and comparing it with that of MYULA. Note that the restoration results of the proposed approach are close to the ones given by MYULA although the latter targets another posterior distribution. We would like to emphasize that although SGS targets an approximate probability distribution, the approximation is controlled with a single parameter ρ that can be made arbitrarily small. On the contrary, MYULA suffers here from a lot of approximations namely the absence of accept/reject step, the use of the Anscombe VST and the Douglas-Rachford splitting scheme to compute the proximity operator. Although the first approximation can be controlled with a single parameter, the second one is not justified in all scenarios and the last one implies an additional computational cost since it is iterative.

method	M	MAE	norm. MAE
SGS	30	6.82	0.07
MYULA	30	5.93	0.06
SGS	100	6.41	0.06
MYULA	100	6.12	0.06

Table 2.6: Performance results associated to the MMSE.

Efficiency of SGS – As before, we assess the efficiency of the proposed sampler by computing the ESSR with samples obtained during the stationary regime. In this experiment, we use the slowest component of $\boldsymbol{\theta}$, that is the one having the largest variance, as scalar summary to compute this ESSR. Table 2.7 depicts its value along with the CPU time required to obtain one sample from SGS. For information, we also indicate the ESSR associated to MYULA although the latter is not directly comparable to that of SGS since the two samplers target two different distributions. One can denote that the mixing properties of SGS and MYULA are close which confirms that SGS can be effectively used to solve similar Poissonian image restoration problems instead of MYULA.

method	ESSR	T_1 [seconds]
SGS	1.58	0.032
MYULA	1.65	0.048

Table 2.7: Efficiency of SGS in terms of ESSR. Although MYULA targets a different probability distribution, we also indicate its ESSR for sake of information and comparison.

2.4 Conclusion

This chapter presented the so-called split Gibbs sampler (SGS) which targets the approximate distribution induced by an AXDA model. We showed

that this sampler shares strong relations with common optimization methods (e.g., quadratic penalty methods and ADMM) and benefits from their main advantages. We can list for instance their divide-to-conquer feature which lead to simpler inference steps or their scalability in distributed architectures. Apart from these nice properties, SGS has been shown to compete with and even improve upon state-of-the-art approaches on two challenging and high-dimensional image processing problems. The proposed sampler also allowed to tackle difficult sampling problems beyond Gaussian likelihood functions . We indeed showed that SGS can be used to tackle efficiently Poissonian image restoration with synthesis sparsity prior distributions. This last application paved the way toward efficient fully Bayesian approaches for even more complicated models (e.g., Poisson-Gaussian noise) or richer models using sophisticated regularization functions (e.g. total generalized variation). In the next chapter, we will complement these experimental results by assessing theoretically the mixing properties of SGS and showing that the latter compete with that of efficient MCMC sampling methods.

3

A non-asymptotic convergence analysis of the Split Gibbs sampler

"I may see a mountain where there is only an anthill."

— Agatha Christie, *The A.B.C. Murders*

Chapter contents

Chapter 2 focused on a specific Monte Carlo sampling algorithm, called split Gibbs sampler (SGS), based on the AXDA framework introduced in Chapter 1. We showed in Section 2.3 that SGS appeared to provide empirically state-of-the-art performance on challenging Bayesian inference problems. Nevertheless, its theoretical behavior in high dimension is currently unknown. In this chapter, we propose a detailed convergence analysis of SGS. We put the emphasis on non-asymptotic, simple and computable bounds in order to provide practitioners with a turn-key algorithm. In addition, we pay attention to the scaling of these bounds with the dimension of the problem, the prescribed precision and the condition number of the potential function of the target posterior density.

In Section 3.1, we give a short introduction to mixing time, that is the minimum number of steps of a Markov chain such that the distance to stationarity is small. Then under regularity conditions, we establish in Section 3.2 explicit convergence rates for SGS using Ricci curvature and coupling ideas. Combining these results to our bounds on the bias of AXDA models derived in Section 1.3, we provide mixing time bounds with explicit dependencies w.r.t. the dimension of the problem, its associated condition number and the prescribed precision. We support our theory with numerical illustrations in Section 3.3. Proofs are collected in Appendix C at the end of the manuscript.

The results of this chapter are the consequences of an international collaboration partly conducted during a visiting period at the Department of Statistics of the University of Oxford between February and May 2019. This work has been submitted to an international journal:

M. Vono, D. Paulin, and A. Doucet (2019). "Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting." In 2nd round of review, *Journal of Machine Learning Research*. arXiv: 1905.11937

3.1 Markov chains and mixing

Whenever a MCMC algorithm is used, an important question to answer is: How long should I wait before obtaining a sample with a distribution close to the target one π? More formally, if we consider a particular MCMC algorithm defining a Markov chain with initial distribution ν, Markov transition kernel $\mathbf{P}$ and invariant target distribution π, this means that we would like to estimate the minimum number of iterations t of the algorithm such that

$$D(\nu\mathbf{P}^t, \pi) \leq \epsilon, \tag{3.1}$$

where D stands for a given statistical distance between two probability measures and $\epsilon > 0$ is the prescribed precision. As such, we define the ϵ-mixing time associated to a statistical distance D and the precision $\epsilon > 0$ as

$$t_{\text{mix}}(\epsilon; \nu) = \min \left\{ t \geq 0 \;\middle|\; D(\nu\mathbf{P}^t, \pi) \leq \epsilon \right\}, \tag{3.2}$$

which stands for the minimum number of steps of the Markov chain such that its distribution is at most at an ϵ D-distance from the invariant target distribution π.

For $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\mathcal{A} \in \mathcal{B}(\mathbb{R}^d)$, the Markov transition kernel $\mathbf{P}$ is defined by $\mathbf{P}(\boldsymbol{\theta}, \mathcal{A}) = \Pr\left(\boldsymbol{\theta}^{(t+1)} \in \mathcal{A} | \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}\right)$ and is such that $\mathbf{P}(\boldsymbol{\theta}, \cdot)$ defines a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ for all $\theta \in \mathbb{R}^d$ and $\mathbf{P}(\cdot, \mathcal{A})$ is $\mathcal{B}(\mathbb{R}^d)$-measurable for all $\mathcal{A} \in \mathcal{B}(\mathbb{R}^d)$.

Answering precisely this question is a difficult task in general which highly depends on the MCMC algorithm, properties of the target distribution, and the distance D used to assess the convergence of the marginal distribution of the Markov chain to π (e.g., Wasserstein or total variation distances). As such, a vast literature dedicated to the convergence of MCMC algorithms is available in various scenarios. Due to the huge number of contributions, we cannot review all of them here. Instead, we give in this section a brief account of the main theoretical research routes that have been considered to answer convergence properties of MCMC approaches.

Several results have been dedicated first to finding qualitative bounds on (3.2) by focusing on the rate of convergence of MCMC algorithms instead of deriving explicit expressions for the mixing time bound (Nummelin and Tuominen, 1983; Tuominen and Tweedie, 1994). In short, these results built upon the analysis of the return time of a Markov chain towards a so-called *atom* (i.e., an accessible set) and the Nummelin's splitting decomposition (Stone and Wainger, 1967; Nummelin, 1978) in order to qualify the rate of convergence of the marginal law of the Markov chain towards π. As an example, given a non-decreasing rate function $r : \mathbb{R}_+ \to \mathbb{R}$ (e.g., $r(t) = t^\beta$ with $\beta > 0$), Nummelin and Tuominen (1983) derived conditions upon which

$$\lim_{t \to \infty} r(t) \left\| \nu\mathbf{P}^t - \pi \right\|_{\text{TV}} = 0. \tag{3.3}$$

Albeit helpful to ensure geometric or sub-geometric convergence of Markov chains, these results cannot be used to answer precisely the question raised at the beginning of this section.

To this purpose, another line of research instead focused on deriving non-asymptotic and computable bounds, that is bounds which can be explicitly determined from the Markov transition kernel $\mathbf{P}$. Such bounds

usually admit the following form:

$$r(t) \left\| \nu \mathbf{P}^t - \pi \right\|_{\mathrm{TV}} \leq C(\nu, \mathbf{P}), \tag{3.4}$$

where $C(\nu, \mathbf{P})$ stands for a constant depending on the initial distribution ν and the Markov kernel $\mathbf{P}$. The general expression of this constant can be determined via so-called drift and minorization conditions, see for instance the works by Meyn and Tweedie (1993), Jarner and Hansen (2000), Roberts and Stramer (2002), Fort and Moulines (2003), and Roberts and Rosenthal (2004). Nonetheless, although such conditions can be easily verified in simple scenarios (e.g., bivariate normal model), their derivation becomes challenging for complicated target distributions (Rosenthal, 1995; Jones and Hobert, 2001).

Following the path of recent works (Choi and Hobert, 2013; Durmus and Moulines, 2017; Dalalyan and Karagulyan, 2019), we focus in this chapter on deriving user-friendly mixing time bounds, that is bounds that are explicit, general and derived under conditions that can be easily checked in practice. These results are presented in the following section.

3.2 Explicit mixing time bounds

In this section, we state our results concerning non-asymptotic bounds on the mixing time of SGS detailed in Algorithm 1 in the "full splitting" scenario that is when $p = b$ in (2.2). This means that the joint approximate density $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ reads

$$\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b}) \propto \prod_{i=1}^{b} \exp\left(-f_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{A}_i \boldsymbol{\theta}\|^2}{2\rho^2} \right). \tag{3.5}$$

In the sequel, we will use the notation f to refer to the potential of π, that is for all $\boldsymbol{\theta} \in \mathbb{R}^d$, $f(\boldsymbol{\theta}) = \sum_{i=1}^{b} f_i(\mathbf{A}_i \boldsymbol{\theta})$.

3.2.1 *Assumptions*

To prove these non-asymptotic results, we shall introduce various regularity conditions (not used all at once) listed in Assumption 1.

Assumption 1 (General assumptions)**.**

- (A_0) *For every* $j \in [b]$, $\inf_{\mathbf{z}_j \in \mathbb{R}^{d_j}} f_j(\mathbf{z}_j) > -\infty$ *(f_j are bounded from below), and for at least one* $i \in [b]$ *we have* $d_i = d$, $\mathbf{A}_i$ *is full rank, and* $\exp(-f_i(\mathbf{z}_i))$ *integrable on* $\mathbb{R}^d$.
- (A_1) f_i *is continuously differentiable and admits a* M_i*-Lipschitz continuous gradient, i.e.,* $\exists\ M_i \geq 0$ *such that* $\left\| \nabla f_i(\mathbf{z}_i') - \nabla f_i(\mathbf{z}_i) \right\| \leq M_i \left\| \mathbf{z}_i' - \mathbf{z}_i \right\|$.
- (A_2) f_i *is convex, i.e. for every* $\alpha \in [0,1]$, $\mathbf{z}_i, \mathbf{z}_i' \in \mathbb{R}^{d_i}$, *we have* $f_i(\alpha \mathbf{z}_i + (1-\alpha)\mathbf{z}_i') \leq \alpha f_i(\mathbf{z}_i) + (1-\alpha) f_i(\mathbf{z}_i')$.

(A_3) *f_i is m_i-strongly convex, i.e., $\exists\ m_i > 0$ such that $f_i(\mathbf{z}_i) - \frac{m_i\|\mathbf{z}_i\|^2}{2}$ is convex.*

(A_4) *$d_1 = \ldots = d_b = d$ and $\mathbf{A}_1 = \ldots = \mathbf{A}_b = \mathbf{I}_d$.*

Assumptions (A_1) and (A_2) on the individual potential functions f_i stand for regularity assumptions which are standard in the optimization literature (Beck, 2017). Some of them have also been recently used in the statistical community to derive non-asymptotic mixing time bounds, see for instance the works by Durmus, Majewski, and Miasojedow (2019) and Dalalyan (2017).

3.2.2 *Dimension-free convergence rates*

We first prove a key result related to the Ricci curvature of SGS which allows us to derive explicit convergence rates for this algorithm.

Lower bound on the Ricci curvature of the SGS kernel – The SGS sampler described in Algorithm 1 generates a Markov chain $\{\boldsymbol{\theta}^{(t)}\}_{t\geq 1}$ of transition kernel $\mathbf{P}_{\mathrm{SGS}}$ defined by

$$\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int_{\mathbf{z}_{1:b}} \pi_\rho(\mathbf{z}_{1:b}|\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta}'|\mathbf{z}_{1:b})\mathrm{d}\mathbf{z}_{1:b}, \tag{3.6}$$

where the conditional distributions associated to π_ρ are defined in (2.3) and (2.4). For any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \mathbb{R}^d$, given a metric $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, the coarse Ricci curvature $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ of $\mathbf{P}_{\mathrm{SGS}}$, introduced by Ollivier (2009), equals

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1 - \frac{W_1(\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \cdot), \mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}', \cdot))}{w(\boldsymbol{\theta}, \boldsymbol{\theta}')}, \tag{3.7}$$

for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \mathbb{R}^d$. We can also define this quantity for p-Wasserstein distances for any $1 \leq p \leq \infty$ by

$$K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1 - \frac{W_p(\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \cdot), \mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}', \cdot))}{w(\boldsymbol{\theta}, \boldsymbol{\theta}')}. \tag{3.8}$$

In Theorem 3, we show under Assumption (A_3) – strong convexity of the potential function – that, for any $1\leq p \leq \infty$ and a suitable metric w, $K_p(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is lower bounded by a simple quantity having an explicit dependence w.r.t. the tolerance parameter ρ and the strong convexity constants of the potential functions $f_{i,i\in[b]}$.

Theorem 3. *Suppose that (A_0) and (A_3) hold. Define the metric*

$$w(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left\| \left(\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i \right)^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}') \right\|. \tag{3.9}$$

Let

$$K_{\mathrm{SGS}} := 1 - \left\| \left(\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i \right)^{-1/2} \left(\sum_{i=1}^{b} \frac{\mathbf{A}_i^T \mathbf{A}_i}{1 + m_i \rho^2} \right) \left(\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i \right)^{-1/2} \right\|. \tag{3.10}$$

Then for the transition kernel $\mathbf{P}_{\mathrm{SGS}}$ *of SGS,* $K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq K_{\mathrm{SGS}}$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \mathbb{R}^d$*, any* $1 \leq p \leq \infty$*.*

Proof. The proof is postponed to Appendix C.1. **Disclaimer:** This proof has been derived by Daniel Paulin and not by the author of this manuscript. □

As shown in the following corollary, Theorem 3 implies that the convergence rate of SGS towards its invariant distribution is governed by the constant K_{SGS} defined in (3.10).

Corollary 4. *Suppose that* (A_0) *and* (A_3) *hold. Then, for any* $1 \leq p \leq \infty$*, any initial distribution* ν *on* $\mathbb{R}^d$*, we have*

$$W_p^w(\nu \mathbf{P}_{\mathrm{SGS}}^t, \pi_\rho) \leq W_p^w(\nu, \pi_\rho) \cdot (1 - K_{\mathrm{SGS}})^t, \tag{3.11}$$

$$\|\nu \mathbf{P}_{\mathrm{SGS}}^t - \pi_\rho\|_{\mathrm{TV}} \leq \mathrm{Var}_{\pi_\rho}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho} \right) \cdot (1 - K_{\mathrm{SGS}})^t, \tag{3.12}$$

where W_p^w *denotes Wasserstein distance of order* p *w.r.t. the metric* w *defined in* (3.9).

Proof. The proof is postponed to Appendix C.2. **Disclaimer:** This proof has been derived by Daniel Paulin and not by the author of this manuscript. □

An attractive property of the convergence rate K_{SGS} is that it is dimension-free, only depends on b, ρ and the strong convexity parameter m_i, and neither requires differentiability nor smoothness of the potential functions $(f_i; i \in [b])$. This is of interest since Corollary 4 can be applied to many problems where non-differentiable potential functions are considered; see e.g., the works by Li and Lin (2010), Xu and Ghosh (2015), and Gu et al. (2014).

Illustrations on the toy Gaussian example – Before proving our mixing time bounds for the SGS, we perform a simple sanity check on a toy Gaussian example in order to assess the tightness of our convergence bounds stated in Corollary 4. The target distribution is chosen as a scalar Gaussian

$$\pi(\theta) = \mathcal{N}\left(\theta; \mu, \frac{\sigma^2}{b}\right), \tag{3.13}$$

where $b \geq 1$ and $\sigma > 0$. In the sequel, we set $\mu = 0$, $\sigma = 3$ and $b = 10$ and consider two different and complementary splitting strategies.

For $p \geq 1$, and a metric $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the Wasserstein distance of order p between two probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is defined by

$$W_p^w(\mu, \nu) = \left(\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d} w(\boldsymbol{\theta}, \boldsymbol{\theta}')^p \mathrm{d}\pi(\boldsymbol{\theta}, \boldsymbol{\theta}') \right)^{1/p}$$

where $\mathcal{U}(\mu, \nu)$ is the set of all probability measures which admit μ and ν as marginals. For $p = \infty$, the Wasserstein distance of order ∞ is defined as

$$W_p^w(\mu, \nu) = \inf_{\pi \in \mathcal{U}(\mu,\nu), (\theta,\theta') \sim \pi} \operatorname{ess\,sup} w(\theta, \theta').$$

In the case when w is the Euclidean metric, we will denote these by $W_p(\mu, \nu)$.

Splitting strategy 1. We set $f_i(\theta) = (2\sigma^2)^{-1}(\theta - \mu)^2$ for all $i \in [b]$. The marginal of θ under the joint (3.5) admits the closed-form expression

$$\pi_\rho(\theta) = \mathcal{N}\left(\theta; \mu, \frac{\sigma^2 + \rho^2}{b}\right).$$

Splitting strategy 2. We set $f(\theta) := f_1(\theta) = b(2\sigma^2)^{-1}(\theta - \mu)^2$. This yields

$$\pi_\rho(\theta) = \mathcal{N}\left(\theta; \mu, \frac{\sigma^2}{b} + \rho^2\right).$$

In this case, the θ-chain follows a Gaussian auto-regressive process of order 1. We can thus compute analytically the Markov transition kernel $\nu \mathbf{P}_{\mathrm{SGS}}^t$ and the total variation and 1-Wasserstein distances between this kernel and the invariant distribution π_ρ; see Appendix C.6 for details. For this toy Gaussian example, the convergence rate of SGS is governed by

$$K_{\mathrm{SGS}} = \frac{\rho^2}{\sigma^2 + \rho^2}, \text{ for the splitting strategy 1,} \tag{3.14}$$

$$K_{\mathrm{SGS}} = \frac{b\rho^2}{\sigma^2 + b\rho^2}, \text{ for the splitting strategy 2.} \tag{3.15}$$

Figure 3.1 illustrates our convergence bounds for each splitting strategy and associated statistical distance. For the total variation case, the slope in log-scale associated to our bound, which equals $\log(1 - K_{\mathrm{SGS}})$, appears to be sharp since it matches the slope associated to the observed convergence rate. Regarding the Wasserstein scenario, the slope associated to our bound is roughly equal to twice the real slope in log-scale, and hence is a bit conservative.

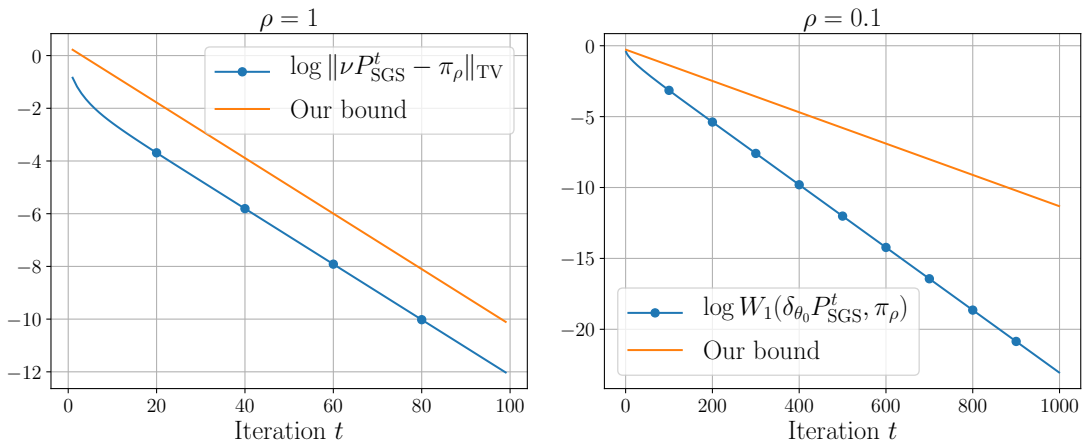


Figure 3.1: From left to right: $\left\|\nu P_{\mathrm{SGS}}^t - \pi_\rho\right\|_{\mathrm{TV}}$ with $\nu(\theta) = \mathcal{N}(\theta; \mu, \sigma^2/b)$ and $W_1(\delta_{\theta_0} P_{\mathrm{SGS}}^t, \pi_\rho)$ with $\theta_0 = 0$ along with the bounds shown in Theorem 3 for the toy Gaussian model (3.13).

We are now ready to prove our main results, namely mixing time bounds associated to SGS for an initial density π which is smooth and strongly log-concave. These assumptions will be weakened in Section 3.2.4.

3.2.3 *User-friendly mixing time bounds*

We consider two cases depending on the statistical distance of interest: Wasserstein distance or total variation distance. In both cases, we will derive explicit expressions for the mixing time of SGS and will compare them to the ones recently obtained in the MCMC literature. For the moment, these bounds consider the single splitting scenario where $b = 1$ and $f(\boldsymbol{\theta}) := f_1(\mathbf{A}_1\boldsymbol{\theta})$. Although the multiple splitting scenario is the most interesting one in practice, note that restricting ourselves to the single splitting case still permits to have a first understanding of the scaling of SGS in high-dimensional and possibly ill-conditioned problems. We are currently working on the generalization of these results in the multiple splitting scenario involving operators $(\mathbf{A}_i; i \in [b])$ acting on $\boldsymbol{\theta}$.

1-Wasserstein distance – We begin by considering the case where the statistical distance of interest is the 1-Wasserstein distance. By combining our error bounds on $W_1(\pi, \pi_\rho)$ from Proposition 7 to the convergence bound (3.11) in Corollary 4, we obtain the following complexity result.

Theorem 4. *Suppose that $b = 1$ and that Assumptions (A_0), (A_1), (A_3) and (A_4) hold. Let $\boldsymbol{\theta}^*$ be the unique minimizer of f_1 and let $\nu = \delta_{\boldsymbol{\theta}^*}$ be the initial distribution. Suppose that $\epsilon \leq 1$. Then, with the choice*

$$\rho^2 = \max\left(\frac{\epsilon^2}{4m_1}, \frac{\epsilon}{\sqrt{m_1 M_1}}\right), \tag{3.16}$$

and number of iterations $t \geq t_{\mathrm{mix}}(\epsilon\sqrt{d/m_1}; \nu)$ where

$$t_{\mathrm{mix}}(\epsilon\sqrt{d/m_1}; \nu) = \frac{\log\left(\frac{3}{\epsilon}\right)}{\log\left(1 + \max\left(\frac{\epsilon^2}{4}, \epsilon\sqrt{\frac{m_1}{M_1}}\right)\right)}, \tag{3.17}$$

we have

$$W_1(\nu P_{\mathrm{SGS}}^t, \pi) \leq \frac{\epsilon}{\sqrt{m_1}}\sqrt{d}. \tag{3.18}$$

This implies that, using t steps of SGS, we can obtain a sample that has a Wasserstein distance from the target π at most equal to $\frac{\epsilon\sqrt{d}}{\sqrt{m_1}}$.

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

Several comments can be made on the result stated in Theorem 4. The expressions of both the choice of the tolerance parameter (3.16) and the mixing time (3.17) are simple and can be computed in practice. These nice properties along with the explicit dependencies of the mixing time of SGS w.r.t. the condition number $\kappa := M_1/m_1$ of f and the desired precision ϵ make Theorem 4 of particular interest for practitioners. In addition, under smoothness and strong convexity of the potential f (see Assumption 1), one can show that $W_1(\delta_{\boldsymbol{\theta}^*}, \pi) \leq \sqrt{d/m_1}$ (Durmus and Moulines, 2019, Proposition 1). This quantity can be interpreted as the typical deviation associated to the sampling problem. Under the assumptions of Theorem

4, it follows that $W_1(\nu P_{\mathrm{SGS}}^t, \pi)$ is upper bounded by ϵ times this typical deviation. Note that considering the relative precision $\epsilon\sqrt{d/m_1}$ yields a mixing time bound (3.17) which is invariant to the scaling of f (that is replacing f by αf with $\alpha > 0$).

For a fixed condition number κ and a sufficiently small precision ϵ, (3.17) implies that the mixing time of SGS scales as $\mathcal{O}(\sqrt{\kappa}\epsilon^{-1}\log(3\epsilon^{-1}))$. To be competitive with other MCMC algorithms, such as those based on Langevin or Hamiltonian dynamics, we have to ensure that the auxiliary variable $\mathbf{z}_1$ can be efficiently drawn at each iteration of Algorithm 1. In Corollary 7 in Appendix C.4, we establish that this is possible by showing that if $\epsilon \leq 1/(d\sqrt{\kappa})$, then sampling $\mathbf{z}_1$ given $\boldsymbol{\theta}$ can be performed by rejection sampling with $\mathcal{O}(1)$ expected evaluations of f and its gradient. Based on this rejection sampling scheme, Table 3.1 compares our complexity result for SGS with single splitting with the ones derived recently in the literature. It shows that SGS compares favourably to competing methods when $0 < \epsilon \leq 1/(d\sqrt{\kappa})$.

Reference	Method	Validity	Evals
Durmus, Majewski, and Miasojedow (2019)	Unadjusted Langevin	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa}{\epsilon^2}\right)$
Cheng et al. (2018)	Underdamped Langevin	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa^2}{\epsilon}\right)$
Dalalyan and Riou-Durand (2018)	Underdamped Langevin	$0 < \epsilon \leq \frac{1}{\sqrt{\kappa}}$	$\mathcal{O}^*\left(\frac{\kappa^{3/2}}{\epsilon}\right)$
Chen and Vempala (2019)	Hamiltonian Dynamics	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa^{3/2}}{\epsilon}\right)$
this work	SGS with single splitting	$0 < \epsilon \leq \frac{1}{d\sqrt{\kappa}}$	$\mathcal{O}^*\left(\frac{\kappa^{1/2}}{\epsilon}\right)$

Table 3.1: Comparison of convergence rates in Wasserstein distance with the literature, starting from the minimizer $\boldsymbol{\theta}^*$ of the m_1-strongly convex and M_1-smooth potential $f_1(\boldsymbol{\theta})$, with condition number $\kappa = M_1/m_1$. SGS with single splitting is implemented based on rejection sampling. $\mathcal{O}^*(\cdot)$ denotes $\mathcal{O}(\cdot)$ up to polylogarithmic factors. In the last column, the complexity stands for the number of gradient and function evaluations to get a W_1 error of $\frac{\epsilon\sqrt{d}}{\sqrt{m_1}}$.

Total variation distance – In this section, we consider the case where one is interested in mixing time bounds w.r.t. the total variation distance. For this scenario, the following theorem states explicit mixing time bounds.

Theorem 5. *Suppose that Assumptions (A_0), (A_1) and (A_3) hold. Assume that $b = 1$, $d_1 = d$ and $\mathbf{A}_1$ has full rank. Let $\boldsymbol{\theta}^*$ be the unique minimizer of $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) = f_1(\mathbf{A}_1\boldsymbol{\theta})$. Let $\nu(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, (M_1\mathbf{A}_1^T\mathbf{A}_1)^{-1})$ be the initial distribution. Then for any $0 < \epsilon \leq 1$, with the choice*

$$\rho^2 \leq \frac{\epsilon}{dM_1}, \tag{3.19}$$

and number of iterations $t \geq t_{\mathrm{mix}}(\epsilon; \nu)$ where

$$t_{\mathrm{mix}}(\epsilon; \nu) = \frac{\log\left(\frac{2}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}}, \tag{3.20}$$

for $K_{\mathrm{SGS}} = \frac{m_1\rho^2}{1+m_1\rho^2}$ and

$$C = \frac{5d}{8} + \frac{d}{2}\log\left(\frac{M_1}{m_1}\right),$$

we have

$$\|\nu P_{\mathrm{SGS}}^t - \pi\|_{\mathrm{TV}} \leq \epsilon.$$

This means that starting from ν, after t step of SGS, we are at a TV-distance at most ϵ from π.

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

Again, note that the bounds for both the tolerance parameter (3.19) and the ϵ-mixing time (3.20) are explicit and can be computed in practice.

Reference	Method	Validity	Evals
Durmus and Moulines (2017)	ULA, $\nu = \delta_{\boldsymbol{\theta}^*}$	$0 \leq \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d/\epsilon^2\right)$
Cheng and Bartlett (2018); Durmus, Majewski, and Miasojedow (2019)	ULA, $\nu = \nu_m$	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d/\epsilon^2\right)$
Dalalyan (2017)	ULA, $\nu = \nu_M$	$0 \leq \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d^3/\epsilon^2\right)$
Dwivedi et al. (2019)	MALA, $\nu = \nu_M$	$0 < \epsilon \leq 1$	$\mathcal{O}\left(\kappa^2 d^2 \log^{1.5}\left(\frac{\kappa}{\epsilon^{1/d}}\right)\right)$
this work	SGS, $\nu = \nu_M$	$0 < \epsilon \leq 1$	$\mathcal{O}^*(\kappa d^2/\epsilon)$

Table 3.2: Comparison of convergence rates in TV distance with the literature, starting from a Gaussian distribution centered at the minimizer $\boldsymbol{\theta}^*$ of the m_1-strongly convex and M_1-smooth potential $f_1(\mathbf{A}_1\boldsymbol{\theta})$, with condition number $\kappa = \frac{M_1}{m_1}$. SGS is implemented based on rejection sampling. $\mathcal{O}^*(\cdot)$ denotes $\mathcal{O}(\cdot)$ up to polylogarithmic factors, $\nu_m(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, \frac{\mathbf{I}_d}{m_1})$ and $\nu_M(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, \frac{\mathbf{I}_d}{M_1})$. The notation ν stands for the initialization of each method.

If we denote the condition number of the potential f by $\kappa := M_1/m_1$, this theorem implies that $t_{\mathrm{mix}}(\epsilon; \nu)$ scales as $\mathcal{O}(d^2\kappa/\epsilon)$ up to polylogarithmic factors. In this scenario, Table 3.2 compares our complexity results for SGS implemented using rejection sampling with existing results in the literature. For the same initialization ν, we have better dependencies than ULA w.r.t. both κ, d and ϵ. However, MALA seems to have better convergence rates in total variation distance in general, except in badly conditioned situations, where the rates for SGS can be better.

3.2.4 *Nonstrongly log-concave target density*

The complexity results shown in Section 3.2.3 assumed the potential f_1 is strongly convex which might be restrictive. In this section, we extend our explicit mixing time bound for the total variation distance to densities which are smooth (see Assumption (A_1)) but such that f_1 only satisfies the standard convexity assumption (A_2) instead of satisfying the strong convexity assumption (A_3).

Similarly to Dalalyan (2017) and Dwivedi et al. (2019), we will weaken our strongly-convex assumption (A_3) by approximating the potential f_1 with a strongly convex one and then applying our previous proof techniques to this approximation. More precisely, instead of the initial target density π, we now consider the approximate density $\tilde{\pi}(\boldsymbol{\theta}) \propto \exp(-\tilde{f}(\boldsymbol{\theta}))$ with

$$\tilde{f}(\boldsymbol{\theta}) = f_1(\mathbf{A}_1\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2, \tag{3.21}$$

where $\lambda > 0$ and $\boldsymbol{\theta}^*$ stands for the minimizer of f. For the multiple splitting strategy studied in the previous section, this approximation leads us to consider an additional error term to bound, namely $\|\pi - \tilde{\pi}\|_{\mathrm{TV}}$. If

$\int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^4 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq d^2 R^2$ with $R > 0$, then with the choice $\lambda = 4\epsilon/(3dR)$, we have $\|\pi - \tilde{\pi}\|_{\mathrm{TV}} \leq \epsilon/3$. It follows that applying Theorem 5 with the new smooth and strongly-convex constants $\tilde{M}_1 = M_1 + \lambda$ and $\tilde{m}_1 = \lambda$ yields the complexity result stated hereafter.

Corollary 5. *Suppose that $b = 1$, $d_1 = d$, $\mathbf{A}_1$ has full rank, that Assumptions (A_0), (A_1) and (A_2) hold and*

$$\int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^4 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq d^2 R^2$$

for some $R > 0$. Let $\tilde{\pi}$ be defined as in (3.21). Let $\boldsymbol{\theta}^*$ be the unique minimizer of $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) = f_1(\mathbf{A}_1\boldsymbol{\theta})$ and $\tilde{M}_1 = M_1 + \lambda$. Let $\nu(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, (\tilde{M}_1 \mathbf{A}_1^T \mathbf{A}_1)^{-1})$ be the initial distribution. Then for any $0 < \epsilon \leq 1$, with the choices $\lambda = 4\epsilon/(3dR)$ and

$$\rho^2 \leq \frac{2\epsilon}{3d(M_1 + \lambda)},$$

and number of iterations

$$t \geq \frac{\log\left(\frac{3}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}},$$

for

$$K_{\mathrm{SGS}} = \frac{\lambda\rho^2}{1 + \lambda\rho^2} \text{ and } C = \frac{5d}{8} + \frac{d}{2}\log\left(\frac{M_1 + \lambda}{\lambda}\right),$$

we have

$$\|\nu P_{\mathrm{SGS}}^t - \pi\|_{\mathrm{TV}} \leq \epsilon.$$

This means that starting from ν, after t step of SGS applied to the approximate density $\tilde{\pi}$, we are at a TV-distance at most ϵ from π.

Proof. The proof is straightforward. It follows from the triangle inequality and Theorem 5. □

Compared to our mixing time bound derived under the assumption that the potential f_1 is strongly convex, Corollary 5 shows that relaxing the strongly convex assumption affects negatively the dependence w.r.t. both the dimension d and the precision ϵ, as it scales as $\mathcal{O}^*(M_1 d^2/\epsilon^2)$. Nevertheless, this complexity result improves upon that in (Dalalyan, 2017; Dwivedi et al., 2019) for the unadjusted Langevin algorithm (ULA) and the Metropolized random walk (MRW), which respectively scale as $\mathcal{O}^*(M_1^2 d^3/\epsilon^4)$ and $\mathcal{O}^*(M_1^2 d^3/\epsilon^2)$.

3.3 Numerical illustrations

This section aims at illustrating the main theoretical results of Section 3.2. We consider two different examples which satisfy all the assumptions

required in our main statements. The first experiment considers the case where the target π is a multivariate Gaussian density while the second one sets π to be a mixture of two multivariate Gaussian densities. For all approaches and experiments, the initial distribution will be set to $\nu = \mathcal{N}(\boldsymbol{\theta}^*, M^{-1}\mathbf{I}_d)$ for the TV distance and to $\nu = \delta_{\boldsymbol{\theta}^*}$ for the 1-Wasserstein one. The experiments have been carried out on a Dell Latitude 7390 laptop equipped with an Intel(R) Core(TM) i5-8250U 1.60 GHz processor, with 16.0 GB of RAM, running Windows 10.

Recall that M stands for the smoothness parameter of f.

3.3.1 *Multivariate Gaussian density*

In this example, we want to verify empirically the dependencies of the mixing times derived in Section 3.2 w.r.t. the dimension d, the desired precision ϵ and the condition number κ of the potential f. We consider a target zero-mean Gaussian density on $\mathbb{R}^d$

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta}\right), \tag{3.22}$$

where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a positive definite precision matrix. In the sequel, $\mathbf{Q}$ will be chosen to be diagonal and anisotropic, that is $\mathbf{Q} = \mathrm{diag}(q_1, \ldots, q_d)$, with $q_i \neq q_j$ for $i \neq j$. The resulting potential function $f := f_1 = \boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta}/2$ is strongly convex and smooth with parameters $m = \min_{i\in[d]} q_i$ and $M = \max_{i\in[d]} q_i$. Since computing the total variation distance between continuous and multidimensional measures is challenging, we discretized the latter over a set of bins and consider the error between the empirical marginal densities associated to the least favorable direction, that is along the eigenvector associated to m. In the following, we will illustrate our mixing time results for both 1-Wasserstein and total variation distances in the strongly log-concave case.

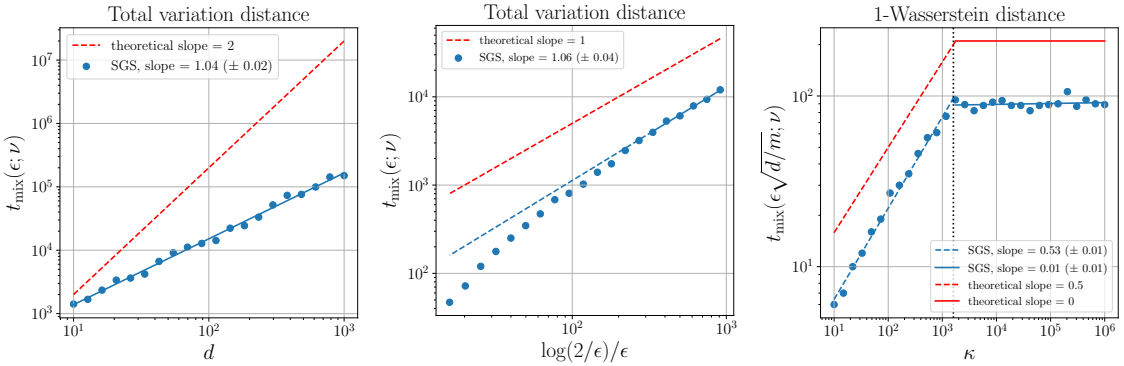


Figure 3.2: Multivariate Gaussian. (left & middle) ϵ-mixing times for the total variation distance and (right) $\epsilon\sqrt{d/m}$-mixing times for the 1-Wasserstein distance.

Dimension dependence. We set here $\epsilon = 0.1$, $m = 1/4$, $M = 1$ such that $\kappa := M/m = 4$ and are interested in the dimension dependence of our ϵ-mixing time result for SGS. We let the dimension d vary between 10^1 and 10^3 and ran SGS for each case. We measured its ϵ-mixing time by recording the smallest iteration such that the discrete total variation error falls below the desired precision ϵ. The mixing time has been averaged over 10 independent runs. Figure 3.2 illustrates the behavior of the mixing

time of SGS w.r.t. the dimension d in log-log scale. In order to assess the dimension dependency, we performed a linear fit and reported the slope of the linear model. According to Table 3.2, the dimension dependence is of order $\mathcal{O}(d^2)$. Interestingly, we found in this example that the dimension dependence of the mixing time of SGS is linear w.r.t. d.

Precision dependence. We set here $d = 2$ and $\kappa = 3$ while the prescribed precision ϵ varies between 6×10^{-3} and 1.6×10^{-1}, and ran SGS for each case. As before, we measured its ϵ-mixing time by recording the smallest iteration such that the discrete total variation error falls below the desired precision ϵ. Figure 3.2 shows the behavior of the mixing time of SGS w.r.t. $\log(2/\epsilon)\epsilon^{-1}$ in log-log scale. For sufficiently small precisions, this figure confirms our theoretical result which states that the mixing time of SGS scales as $\mathcal{O}(\log(2/\epsilon)\epsilon^{-1})$.

Condition number dependence. Regarding the 1-Wasserstein distance and the complexity results depicted in Table 3.1, the main difference between existing MCMC approaches is the dependence w.r.t. the condition number κ of the potential function f. Here, we aim at verifying the latter numerically. To this purpose, we set $d = 10$, $\epsilon = 0.1$ and let κ vary between 10^1 and 10^6. From (3.17), it appears that the dependence of the mixing time of SGS depends on $\max\{\epsilon^2/4, \epsilon/\sqrt{\kappa}\}$. This quantity equals $\epsilon/\sqrt{\kappa}$ for $\kappa \leq 1,600$ and $\epsilon^2/4$ otherwise. Hence, we are expecting to retrieve a dependence in $\kappa^{1/2}$ for small and moderate κ and a mixing time only depending on ϵ for larger values of the condition number. We performed 50 independent runs of SGS and stopped them when their empirical Wasserstein error fell below $\epsilon\sqrt{d/m}$. The results are depicted on Figure 3.2 in log-log scale. As before, we did a linear fit to assess the dependency of the mixing time w.r.t. the condition number κ. The slope of the linear model for SGS equals 0.53 for $\kappa \leq 1,600$ (depicted with a black dotted vertical line) which confirms the theoretical dependence of the order $\mathcal{O}(\kappa^{1/2})$. As expected, the mixing time of SGS becomes independent of κ for larger values.

3.3.2 *Gaussian mixture*

The previous section aimed at verifying empirically the dependencies of our mixing time bounds for SGS. In this second experiment, also considered by Dalalyan (2017) and Dwivedi et al. (2019), we show that the values of the tolerance parameter ρ and the mixing time $t_{\text{mix}}(\epsilon; \nu)$ recommended by Theorem 5 indeed yield approximate samples having a distribution close to π. We also verify that the running time required to generate such samples is reasonable, and compare it to the running time of ULA to achieve the same prescribed precision ϵ. To this purpose, let us consider the simple problem of generating samples from a mixture of two Gaussian densities with density π defined, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$\pi(\boldsymbol{\theta}) = \frac{1}{2(2\pi)^{d/2}} \left(\exp\left(-\frac{\|\boldsymbol{\theta} - \mathbf{a}\|^2}{2} \right) + \exp\left(-\frac{\|\boldsymbol{\theta} + \mathbf{a}\|^2}{2} \right) \right)$$
$$\propto \exp\left(-f(\boldsymbol{\theta})\right), \tag{3.23}$$

where

$$f(\boldsymbol{\theta}) = \frac{1}{2}\left\|\boldsymbol{\theta} - \mathbf{a}\right\|^2 - \log\left(1 + e^{-2\boldsymbol{\theta}^T\mathbf{a}}\right), \tag{3.24}$$

and $\mathbf{a} \in \mathbb{R}^d$ is a fixed vector involved in the mean of each Gaussian density. If $\|\mathbf{a}\| < 1$, one can show that f is M-smooth and m-strongly convex with $m = 1 - \|\mathbf{a}\|^2$ and $M = 1$. In the sequel, we choose $\mathbf{a}$ such that $\|\mathbf{a}\| = 1/\sqrt{2}$, which also implies that the global minimizer of f is $\boldsymbol{\theta}^* = \mathbf{0}_d$. Since π admits a finite second order moment, all the assumptions required in Theorem 5 are verified. We now consider a single splitting strategy on f leading to the joint approximate density $\pi_\rho(\boldsymbol{\theta}, \mathbf{z})$ defined in (3.5) with $b = 1$ and $\mathbf{A}_1 = \mathbf{I}_d$. Under this distribution, the marginal density $\pi_\rho(\boldsymbol{\theta})$ writes

$$\pi_\rho(\boldsymbol{\theta}) = \frac{1}{2(2\pi(1+\rho^2))^{d/2}}\left(\exp\left(-\frac{\|\boldsymbol{\theta}-\mathbf{a}\|^2}{2(1+\rho^2)}\right) + \exp\left(-\frac{\|\boldsymbol{\theta}+\mathbf{a}\|^2}{2(1+\rho^2)}\right)\right), \tag{3.25}$$

and simply corresponds to a mixture of the two initial Gaussian densities but with respective variance now inflated by a factor ρ^2. The one-dimensional approximate density $\pi_\rho(u)$ of the projection $u = \mathbf{a}^T\boldsymbol{\theta}/\|\mathbf{a}\|$ is depicted in Figure 3.3 for $d = 60$ and compared to the true target $\pi(u)$.

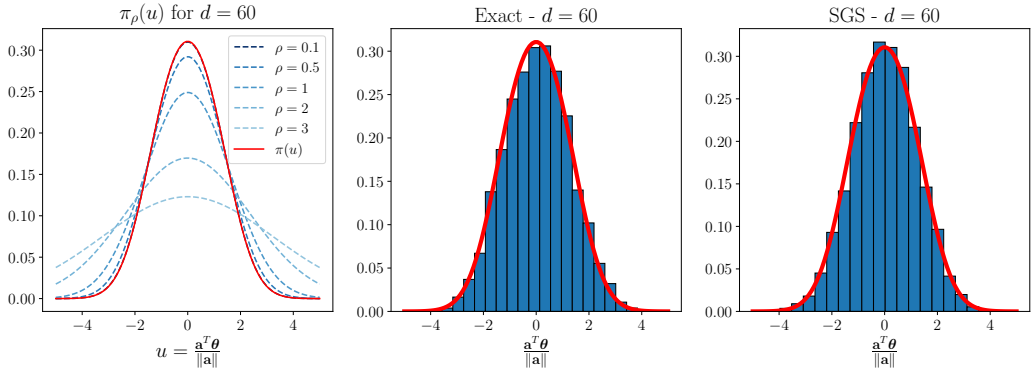


Figure 3.3: Gaussian mixture with $d = 60$. From left to right: behavior of $\pi_\rho(u)$ w.r.t. ρ with $u = \mathbf{a}^T\boldsymbol{\theta}/\|\mathbf{a}\|$; empirical distribution obtained by exact sampling from π; empirical distribution obtained by sampling from π_ρ with the guidelines recommended in Theorem 5. The histograms have been computed using 2500 independent samples and the precision has been set to $\epsilon = 0.1$. In all figures, the red curve stands for $\pi(u)$.

Illustrations of Theorem 5. We now illustrate the guidelines for ρ and the number of iterations t, stated in Theorem 5, to achieve an ϵ-error in total variation distance. To this purpose, we set $\epsilon = 0.1$, $d = 60$ and launched 2500 independent runs of SGS. The conditional distribution of $\mathbf{z}$ given $\boldsymbol{\theta}$ is a mixture of two Gaussians with common covariance matrix $\boldsymbol{\Sigma} = \rho^2/(1+\rho^2)\mathbf{I}_d$, respective mean vectors $\boldsymbol{\mu}_1 = (\boldsymbol{\theta} + \mathbf{a}\rho^2)/(1+\rho^2)$ and $\boldsymbol{\mu}_2 = (\boldsymbol{\theta} - \mathbf{a}\rho^2)/(1+\rho^2)$ and respective weights $w_1 = 1$ and $w_2 = \exp(-4\boldsymbol{\theta}^T\mathbf{a}/(2(1+\rho^2))$. We can sample exactly from this mixture by first drawing a Bernoulli random variable B with probability $p = w_1/(w_1+w_2)$ and then setting $\mathbf{z} = B(\boldsymbol{\xi}+\boldsymbol{\mu}_1) + (1-B)(\boldsymbol{\xi}+\boldsymbol{\mu}_2)$ where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Sigma})$. In order to assess the relevance of the samples generated with SGS, we generated 2500 independent samples directly from π by an exact sampler similar to the one used to sample $\mathbf{z}$. To provide an illustration of the quality of the samples drawn with SGS, we computed the one-dimensional projection $u = \mathbf{a}^T\boldsymbol{\theta}/\|\mathbf{a}\|$ and showed its empirical distribution in Figure 3.3. The empirical distribution of the samples drawn using SGS is indeed

close to π and is visually indistinguishable from the one of the exact samples

Dimension d	4	8	12	16	20	30	40	60
$t_{\text{mix}}(\epsilon;\nu)$ $(\times 10^3)$ for SGS	3	10	23	40	62	138	244	548
$t_{\text{mix}}(\epsilon;\nu)$ $(\times 10^3)$ for ULA	29	87	184	330	532	1,350	2,729	7,742
Efficiency of SGS w.r.t. ULA	10.8	8.6	8.2	8.3	8.6	9.8	11.1	14.1
CPU time [s] for SGS	1	7	29	62	114	335	749	2,416
CPU time [s] for ULA	6	31	135	302	589	1,974	4,766	15,096
Efficiency of SGS w.r.t. ULA	5.6	4.6	4.7	4.9	5.2	5.9	6.4	6.2

Table 3.3: Gaussian mixture. Comparison between SGS and ULA for a prescribed precision $\epsilon = 0.1$. For SGS, $t_{\text{mix}}(\epsilon;\nu)$ has been computed by using Theorem 5 while for ULA, the mixing time bound derived by Dalalyan (2017, Corollary 1) has been used. The CPU time information corresponds to the running time necessary to draw 10^3 independent samples having a distribution at most ϵ total variation distance from π.

Computational complexity of SGS. We now verify empirically the computational complexity of SGS, that is the number of iterations and the overall running time for generating samples with some prescribed precision ϵ. We compare this complexity to that of ULA (Dalalyan, 2017). Starting from the same initial distribution $\nu = \mathcal{N}(\boldsymbol{\theta}^\star, M^{-1}\mathbf{I}_d)$ and with $\epsilon = 0.1$, Table 3.3 reports the number of iterations $t_{\text{mix}}(\epsilon;\nu)$ required, in theory, to obtain a sample whose distribution is at most ϵ in total variation from π and the CPU time needed to generate 10^3 such samples. For ULA, $t_{\text{mix}}(\epsilon;\nu)$ has been computed by using the mixing time bound derived by Dalalyan (2017, Corollary 1). We observe that for $d \in [4, 60]$, both the number of iterations and the running time for generating 10^3 independent samples with SGS are much smaller than that of ULA.

This second experiment confirms our theoretical statement that SGS is able to generate accurate samples for a reasonable computational budget compared to popular alternatives such as ULA.

3.4 Conclusion

As the result of an international collaboration with researchers from Universities of Oxford and Edinburgh, we have provided in this chapter a detailed theoretical study of a promising MCMC algorithm, namely SGS, which is amenable to a distributed implementation and shares strong similarities with quadratic penalty methods. Under a strong log-concavity assumption, we obtained explicit dimension-free convergence rates for this sampler under both Wasserstein and total variation distances. Combined with quantitative bounds on the bias induced by this algorithm, we have derived explicit bounds on its mixing time under reasonable assumptions which can easily be verified in practice. These results showed that SGS can compete and even improve upon standard MCMC schemes in terms of computational complexity. Our theoretical results have been supported with numerical illustrations which confirmed the efficiency of SGS in the single splitting scenario.

In the full splitting strategy, sampling from the approximate joint distribution $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ with SGS involves sampling from the possibly high-

dimensional conditional Gaussian distribution

$$\pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:b}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_{1:b}), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{3.26}$$

with mean vector $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^T \mathbf{A}_i)^{-1} \sum_{i=1}^b \mathbf{A}_i^T \mathbf{z}_i$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \rho^2 (\sum_{i=1}^b \mathbf{A}_i^T \mathbf{A}_i)^{-1}$. In some cases, this Gaussian sampling step can be tackled efficiently. For instance, if the matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is constant across iterations and the dimension d is moderate, its Cholesky decomposition, necessary to sample from (3.26), can be pre-computed in a preliminary step. In addition, for interesting models such as image inpainting or Poissonian image restoration, sampling from (3.26) can be conducted by resorting to the fast Fourier transform, see Section 2.3 in Chapter 2. Nonetheless, this step might become computationally demanding if the matrices $(\mathbf{A}_i; i \in [b])$ yield an arbitrary and high-dimensional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

In Chapter 4, we will propose a unifying approach to tackle such a high-dimensional Gaussian sampling problem by building on a stochastic sampling counterpart of the proximal point algorithm (Rockafellar, 1976).

4

High-dimensional Gaussian sampling: A unifying approach based on a stochastic proximal point algorithm

"Every single bit of information that comes in makes things more difficult."

— Agatha Christie, *Mrs. McGinty's Dead*

Chapter contents

We saw in Chapters 2 and 3 that SGS often involved a potentially high-dimensional Gaussian sampling step. When the dimension of the problem is small or moderate, sampling from this distribution is an old solved problem that raises no particular difficulty. In high-dimensional settings this multivariate sampling task can become computationally demanding so that, even recently, a host of works have focused on the derivation of efficient simulation-based approaches to tackle this problem.

In this chapter, we propose to shed new light on some recently proposed MCMC approaches specifically designed for high-dimensional Gaussian sampling. To this purpose, we build upon a unifying framework which can be interestingly related to the celebrated proximal point algorithm in optimization (Rockafellar, 1976).

In Section 4.1, we present the considered Gaussian sampling problem and its main issues which arise in high-dimensional scenarios. Section 4.2 reviews existing MCMC sampling methods dedicated to Gaussian sampling. Then Section 4.3 introduces the proposed unifying framework and brings together these approaches which come from distinct communities. Possible extensions of the latter and new sampling approaches are also highlighted. Finally, Section 4.4 draws connections between this unifying framework and the proximal point algorithm showing one more time the tight existing bond between optimization and simulation.

This work is based on a wider review paper which has been submitted to an international journal:

M. Vono, N. Dobigeon, and P. Chainais (2020b). "High-dimensional Gaussian sampling: A review and a unifying approach based on a stochastic proximal point algorithm." In 1st round of review, *SIAM Review*. arXiv: 2010.01510

4.1 Problem statement and motivations

This section highlights the considered Gaussian sampling problem, its already-surveyed special instances and its main issues. By recalling these specific instances, this section also defines the focus of this chapter, namely high-dimensional Gaussian sampling with an arbitrary covariance (or precision) matrix.

4.1.1 *Definitions and notation*

We address here the problem of sampling from a d-dimensional Gaussian distribution where d may be large. Its pdf according to the d-dimensional Lebesgue measure, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, writes

$$\pi(\boldsymbol{\theta}) \coloneqq \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}}, \tag{4.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ stand for the mean vector and the covariance matrix of the considered Gaussian distribution, respectively. We assume in the sequel that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite, that is for all $\boldsymbol{\theta} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$, $\boldsymbol{\theta}^T\boldsymbol{\Sigma}\boldsymbol{\theta} > 0$. Hence, its inverse $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$, called *precision* matrix, exists and is also positive definite.

For some approaches and applications, working with the precision $\mathbf{Q}$ rather than with the covariance $\boldsymbol{\Sigma}$ will be more convenient (e.g., for conditional auto-regressive models or hierarchical Bayesian models). In this chapter, we choose to work directly with $\mathbf{Q}$ for the sake of simplicity. When $\mathbf{Q}$ is unknown but $\boldsymbol{\Sigma}$ is available instead, simple and straightforward algebraic manipulations can be used to implement the approaches detailed in the sequel without increasing their computational complexity. Sampling from (4.1) raises several important issues which are mainly related to the structure of $\mathbf{Q}$, that is to the correlations between the d components of $\boldsymbol{\theta}$. In the following paragraphs, we will detail some special instances of (4.1) and well-known associated sampling strategies before focusing on the general Gaussian sampling problem considered in this chapter.

4.1.2 *Usual special instances*

For completeness, this subsection recalls special cases of Gaussian sampling tasks that will not be specifically tackled later but are usual and convenient building blocks. Instead, we point out appropriate references for the interested reader. These special instances include basic univariate sampling and the scenarios where $\mathbf{Q}$ is (i) a diagonal matrix, (ii) a band-matrix or (iii) a circulant matrix. Again, with basic algebraic manipulations, the same samplers can be used when $\boldsymbol{\Sigma}$ has one of these specific structures.

Univariate Gaussian sampling – The most simple Gaussian sampling problem boils down to drawing univariate Gaussian random variables with

mean μ and precision $q > 0$. Generating the latter quickly and with high accuracy has been the topic of much research works in the last 70 years. Such methods can be loosely speaking divided into four groups namely (i) cumulative density function inversion, (ii) transformation, (iii) rejection and (iv) recursive methods; they are now well-documented. Interested readers are invited to refer to the comprehensive overview by Thomas et al. (2007) for more details. For instance, Algorithm 3 details the well-known Box-Muller method which transforms a pair of independent uniform random variables into a pair of Gaussian random variables by exploiting the radial symmetry of the two-dimensional normal distribution.

Algorithm 3: Box-Muller sampler

Input: mean μ and precision $q > 0$.

Draw $u_1,\ u_2 \sim \mathcal{U}((0,1])$.

Set $z_1 = \sqrt{-2\log(u_1)}$.

Set $z_2 = 2\pi u_2$.

Output: $(\theta_1, \theta_2) = \left(\mu + \frac{z_1}{\sqrt{q}}\sin(z_2), \mu + \frac{z_1}{\sqrt{q}}\cos(z_2)\right)$.

Multivariate Gaussian sampling with diagonal covariance matrix – Let us extend the previous sampling problem and now assume that one wants to generate a d-dimensional Gaussian vector $\boldsymbol{\theta}$ with mean $\boldsymbol{\mu}$ and diagonal precision matrix $\mathbf{Q} = \mathrm{diag}(q_1, \cdots, q_d)$. The d components of $\boldsymbol{\theta}$ being independent, this problem is as simple as the univariate one since we can sample the d components in parallel independently. A pseudo-code of the corresponding sampling algorithm is given in Algorithm 4. In this simple scenario, the computational cost required to sample one Gaussian random variable from π is of $\mathcal{O}(d)$ floating point operations (flops).

Algorithm 4: Sampler when $\mathbf{Q}$ is a diagonal matrix

Input: mean vector $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} = \mathrm{diag}(q_1, \cdots, q_d)$.

for $i \in [d]$ **do**

  Draw $\theta_i \sim \mathcal{N}\left(\mu_i, 1/q_i\right)$.

end

Output: $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)^T$.

When $\mathbf{Q}$ is not diagonal, we can no longer sample the d components of $\boldsymbol{\theta}$ independently. However, for well-structured matrices $\mathbf{Q}$, it is still possible to draw the random vector of interest with a reasonable computational cost.

Multivariate Gaussian sampling with sparse or band matrix $\mathbf{Q}$ – A lot of standard Gaussian sampling approaches leverage on the sparsity of the matrix $\mathbf{Q}$. Sparse precision matrices appear for instance when Gaussian Markov random fields (GMRFs) are considered, as illustrated in Figure 4.1. In this figure, German regions are represented graphically where each edge between two regions stands for a common border. These edges can then be described by an adjacency matrix which plays the role of the precision matrix $\mathbf{Q}$ of a GMRF. Since there are few neighbors for each region, $\mathbf{Q}$ is symmetric and sparse. By permuting the rows and columns

of $\mathbf{Q}$, one can build a so-called *band matrix* with minimal bandwidth b where b is the smallest integer $b < d$ such that $Q_{i,j} = 0,\ \forall i > j + b$ (Rue, 2001). Algorithm 5 details the main steps to obtain a Gaussian vector $\boldsymbol{\theta}$ from (4.1) with a band precision matrix $\mathbf{Q}$ at the reduced cost of $\mathcal{O}(b^2 d)$ floating point operations (flops) by using Cholesky's factorization of $\mathbf{Q}$. Similar computational savings can be obtained in the sparse case (Rue and Held, 2005). Note that this method is even simpler when using $\boldsymbol{\Sigma}$: then one uses $\mathbf{C} = \text{chol}(\boldsymbol{\Sigma})$ such that $\boldsymbol{\Sigma} = \mathbf{C}^T\mathbf{C}$ and $\boldsymbol{\theta} = \mathbf{C}\mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. Band matrices naturally appear in specific applications, e.g., when the latter involve finite impulse response linear filters (Idier, 2008). Problems with such structured (sparse or band) matrices have been extensively studied in the literature and as such this chapter will not cover them explicitly. We refer the interested reader to the textbook by Rue and Held (2005) for more details.

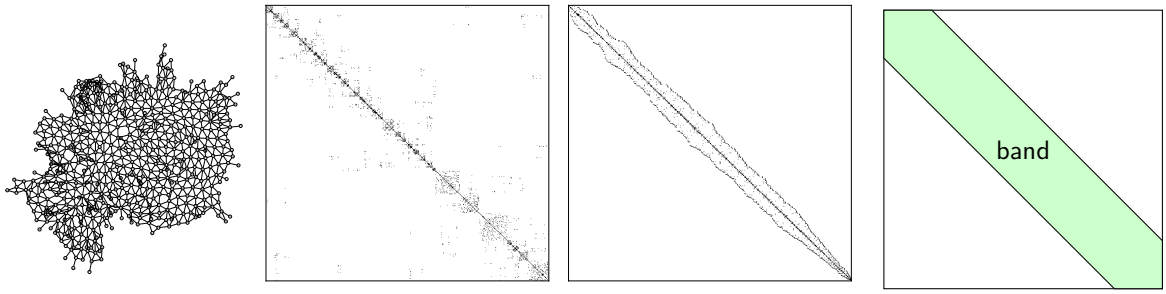


Figure 4.1: From left to right: example of an undirected graph, its associated precision matrix $\mathbf{Q}$ (bandwidth $b = 522$), its re-ordered precision matrix $\mathbf{P}\mathbf{Q}\mathbf{P}^T$ ($b = 43$) where $\mathbf{P}$ is a permutation matrix and a drawing of a band matrix. This graph is defined on the 544 regions of Germany where those sharing a common border are considered as neighbors. The pixels in white are equal to zero.

Algorithm 5: Sampler when $\mathbf{Q}$ is a band matrix

Input: mean vector $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}$.

1 Set $\mathbf{C} = \text{chol}(\mathbf{Q})$. `// Build the Cholesky factor` $\mathbf{C}$ `of` $\mathbf{Q}$, `see Rue and Held (2005, Section 2.4).`

2 Draw $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$.

3 **for** $i \in [d]$ **do** `// Solve` $\mathbf{C}^T\boldsymbol{\theta} = \mathbf{z}$ `by backward substitution.`

4 Set $j = d - i + 1$.

5 Set $m_1 = \min\{j + b, d\}$.

6 Set $\theta_j = \dfrac{1}{C_{j,j}}\left(z_j - \displaystyle\sum_{k=j+1}^{m_1} C_{k,j}\theta_k\right)$.

7 **end**

Output: $\boldsymbol{\mu} + \boldsymbol{\theta}$.

Multivariate Gaussian sampling with block circulant (or Toeplitz) matrix $\mathbf{Q}$ with circulant (or Toeplitz) blocks – An important special case of (4.1) which has already been surveyed (Rue and Held, 2005) is

when $\mathbf{Q}$ is a block circulant matrix with circulant blocks

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 & \dots & \mathbf{Q}_M \\ \mathbf{Q}_M & \mathbf{Q}_1 & \dots & \mathbf{Q}_{M-1} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{Q}_2 & \mathbf{Q}_3 & \dots & \mathbf{Q}_1 \end{pmatrix} \tag{4.2}$$

where $(\mathbf{Q}_i; i \in [M])$ are M circulant matrices. Such structured matrices frequently appear in image processing problems since they translate the convolution operator corresponding to a linear and shift-invariant filters. As an illustration, Figure 4.2 shows the circulant structure of the precision matrix associated with the Gaussian distribution $\pi(\boldsymbol{\theta}) \propto \exp(-\|\mathbf{D}\boldsymbol{\theta}\|^2/2)$. Here, the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ stands for an image reshaped in lexicographic order and $\mathbf{D}$ stands for the Laplacian differential operator with periodic boundaries. In this case the precision matrix $\mathbf{Q} = \mathbf{D}^T\mathbf{D}$ is a circulant matrix (Orieux, Giovannelli, and Rodet, 2010) so that it is diagonalizable in the Fourier domain. Therefore, sampling from (4.1) can be efficiently carried out by using the fast Fourier transform (Wood and Chan, 1994; Dietrich and Newsam, 1997). This approach yields a reduced cost of $\mathcal{O}(d\log(d))$ flops, see Algorithm 6. For Gaussian distributions with more general Toeplitz precision matrices, $\mathbf{Q}$ can be replaced by its circulant approximation and then Algorithm 6 can be used, see Rue and Held (2005) for more details. Although not considered in this chapter, other approaches dedicated to generate stationary Gaussian processes (Li, 2009) have been considered, such as the spectral (Shinozuka and Jan, 1972; Mejía and Rodríguez-Iturbe, 1974) and turning bands (Mantoglou and Wilson, 1982) methods.

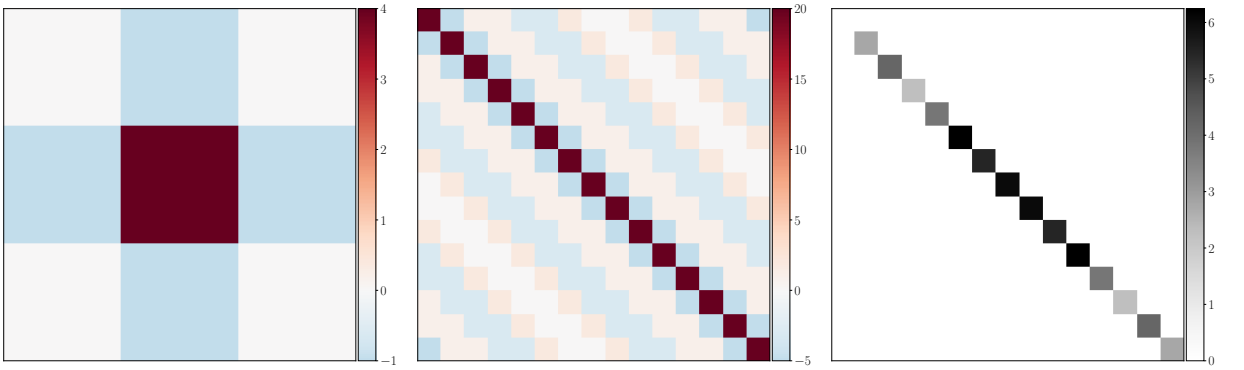


Figure 4.2: From left to right: example of a 3×3 Laplacian filter, the associated circulant precision matrix $\mathbf{Q} = \mathbf{D}^T\mathbf{D}$ when periodic boundary conditions have been considered and its counterpart diagonal matrix $\mathbf{F}\mathbf{Q}\mathbf{F}^H$ in the Fourier domain, where $\mathbf{F}$ and its Hermitian conjugate $\mathbf{F}^H$ are unitary matrices associated with the Fourier and inverse Fourier transforms.

Truncated and intrinsic Gaussian distributions – Eventually, note that several works have focused on sampling from probability distributions closely related to the Gaussian distribution. Two cases are worth being mentioned here, namely the truncated and so-called *intrinsic* Gaussian distributions. The pdf associated to the truncated distributions can be defined as

$$\pi(\boldsymbol{\theta}) = \begin{cases} Z_{\mathcal{D}}^{-1}\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}) & \text{if } \boldsymbol{\theta} \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

Algorithm 6: Sampler when $\mathbf{Q}$ is a block circulant matrix with circulant blocks

Input: M and N, the number of blocks and the size of each block, respectively.

1 Compute $\mathbf{F} = \mathbf{F}_M \otimes \mathbf{F}_N$. `// F_M is the M × M unitary matrix associated to the Fourier transform and ⊗ denotes the tensor product.`

2 Draw $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$.

3 Set $\mathbf{\Lambda} = \mathrm{diag}(\mathbf{q})$. `// q is the d-dimensional vector built by stacking the first columns of each circulant block of Q.`

4 Set $\boldsymbol{\theta} = \mathbf{F}^H \left(\mathbf{F}\boldsymbol{\mu} + \mathbf{\Lambda}^{-1/2}\mathbf{F}\mathbf{z} \right)$.

Output: $\boldsymbol{\theta}$.

where $\mathcal{D} \subset \mathbb{R}^d$ is a subset defined by equalities and/or inequalities, and $Z_{\mathcal{D}}$ is the appropriate normalizing constant. As archetypal examples, truncations on the hypercube are such that $\mathcal{D} = \prod_{i=1}^{d}[a_i, b_i]$, $(a_i, b_i) \in \mathbb{R}^2$, $1 \leq i \leq d$ or $\mathcal{D} = \{\boldsymbol{\theta} \in \mathbb{R}^d | \sum_{i=1}^{d} \theta_d = 1\}$ that limits the domain to the simplex. Sampling algorithms dedicated to these distributions can be found in (Altmann, McLaughlin, and Dobigeon, 2014; Li and Ghosh, 2015; Wilhelm and Manjunath, 2010).

Intrinsic Gaussian distributions are such that $\mathbf{Q}$ is not of full rank, that is $\mathbf{Q}$ may have eigenvalues equal to zero. This yields an improper Gaussian distribution π in (4.1) often used as a prior in GMRFs to remove trend components (Rue and Held, 2005). Sampling from the latter can be done by identifying an appropriate subspace of $\mathbb{R}^d$ where π is proper and then sampling from the proper Gaussian density on this subspace (Besag and Kooperberg, 1995; Parker and Fox, 2012).

All the usual special sampling problems above will not be considered in the following since they have already been exhaustively reviewed and tackled in the literature.

4.1.3 *Problem statement: sampling from a Gaussian distribution with an arbitrary precision matrix* $\mathbf{Q}$

From now on, we will consider the problem of sampling from an *arbitrary non-intrinsic multivariate* Gaussian distribution (4.1), i.e., without assuming any particular structure of the precision or covariance matrix. If $\mathbf{Q}$ is diagonal or well-structured, sampling can be performed efficiently, even in high dimension, see Section 4.1.2 above. When this matrix is unstructured and possibly dense, this is not the case anymore. Then, the main challenges for Gaussian sampling are directly related to handling the precision $\mathbf{Q}$ (or covariance $\mathbf{\Sigma}$) matrix in high dimension. Typical issues include the storage of the full matrix $\mathbf{Q}$ (or $\mathbf{\Sigma}$) and expensive operations such as inversion or square root computation which become prohibitive when d is large. These challenges are illustrated below with an example that typically arises in statistical learning.

Example.

Let us consider a ridge regression problem from a Bayesian perspective (Bishop, 2006). For sake of simplicity and without loss of generality, let assume that the observations $\mathbf{y} \in \mathbb{R}^n$ and the known predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ are such that

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{for } j \in [d]. \tag{4.4}$$

Under these assumptions, we consider the following statistical model associated with the observations $\mathbf{y}$ which writes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{4.5}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. In this example, the standard deviation σ is known and fixed. The conditional prior distribution for $\boldsymbol{\theta}$ is chosen as Gaussian with diagonal covariance matrix, that is

$$\pi(\boldsymbol{\theta}|\tau) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{\theta_i^2}{2\tau}\right), \tag{4.6}$$

$$\pi(\tau) \propto \frac{1}{\tau}, \tag{4.7}$$

where $\tau > 0$ stands for an unknown variance parameter which is given a diffuse and improper (i.e., non-integrable) Jeffrey's prior (Jeffreys, 1946; Robert, 2001). The Bayes' rule then leads to the target joint posterior distribution with density

$$\pi(\boldsymbol{\theta}, \tau|\mathbf{y}) \propto \frac{1}{(2\pi\sigma^2)^{n/2}\tau} \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}{2\sigma^2}\right) \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{\theta_i^2}{2\tau}\right). \tag{4.8}$$

Sampling from this joint posterior distribution can be conducted using a Gibbs sampler (Geman and Geman, 1984; Robert and Casella, 2004) which sequentially samples from the conditional posterior distributions. In particular, the conditional posterior associated to $\boldsymbol{\theta}$ is Gaussian and writes

$$\pi(\boldsymbol{\theta}) \coloneqq \pi(\boldsymbol{\theta}|\mathbf{y}, \tau) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}) \tag{4.9}$$

where

$$\mathbf{Q} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \tau^{-1}\mathbf{I}_d \tag{4.10}$$

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{Q}^{-1}\mathbf{X}^T\mathbf{y}. \tag{4.11}$$

Challenges related to handling the matrix $\mathbf{Q}$ already appear in this classical and simple regression problem. Indeed, $\mathbf{Q}$ is possibly high-

dimensional $(d \times d)$ and dense which potentially rules out its storage. Moreover, the inversion required to compute the mean (4.11) may be very expensive as well. In addition, since τ is unknown, its value changes at each iteration of the Gibbs sampler used to sample from (4.8). Hence, pre-computing the matrix $\mathbf{Q}^{-1}$ once and for all is irrelevant. As an illustration on real data, Figure 4.3 represents three examples of precision matrices $\mathbf{X}^T\mathbf{X}$ for the MNIST (Le Cun et al., 1998), leukemia (Armstrong et al., 2002) and CoEPrA (Ivanciuc, 2006) datasets. One can denote that these precision matrices are potentially both high-dimensional and dense penalizing their numerical inversion at each iteration of the Gibbs sampler. When considering the dataset itself, $\mathbf{X}^T\mathbf{X}$ is usually interpreted as the empirical covariance of the data $\mathbf{X}$. The reader should not be disturbed by the fact that, turning to the variable $\boldsymbol{\theta}$ to infer, $\mathbf{X}^T\mathbf{X}$ will however play the role of a precision matrix.

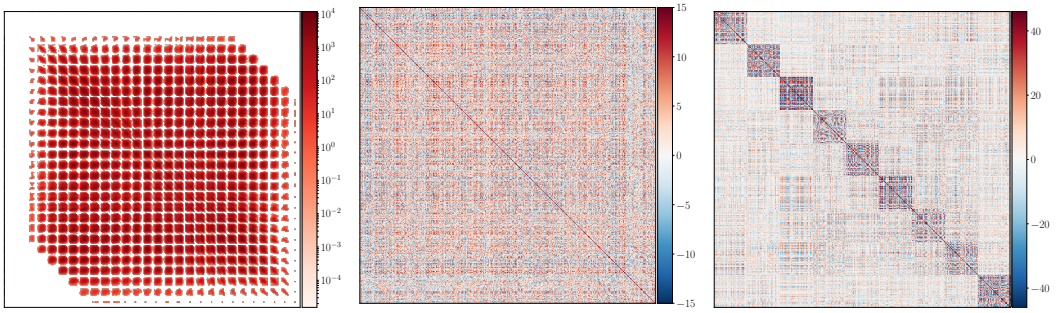


Figure 4.3: Examples of precision matrices $\mathbf{X}^T\mathbf{X}$ for three datasets. Left: MNIST dataset (Le Cun et al., 1998). Only the predictors associated to the digits 5 and 3 have been taken into account for the MNIST dataset (Le Cun et al., 1998). Middle: leukemia dataset (Armstrong et al., 2002). For the leukemia dataset (Armstrong et al., 2002), only the first 5,000 predictors (out of 12,600) have been taken into account. Right: CoEPrA dataset (Ivanciuc, 2006).

Ranging from numerical linear algebra to MCMC methods, hosts of contributions are related to high-dimensional Gaussian sampling; see Vono, Dobigeon, and Chainais (2020b) and references therein for a recent overview. In Section 4.2, we review existing MCMC approaches dedicated to Gaussian sampling before proposing in Section 4.3 a unifying revisit of these samplers by building upon a stochastic sampling counterpart of the celebrated proximal point algorithm (PPA) (Rockafellar, 1976).

4.2 MCMC sampling approaches

In this section, we present a family of sampling approaches, namely MCMC approaches, which build a discrete-time Markov chain $\{\boldsymbol{\theta}^{(t)}\}_{t\in\mathbb{N}}$ having π (or a close approximation of π) as invariant distribution (Robert and Casella, 2004). In the sequel, we state that an approach is exact if the associated MCMC sampler admits an invariant distribution which coincides with π. While being iterative and requiring a reasonable computational cost per iteration, these methods have also been proposed to avoid to work with $\mathbf{Q}$ directly and to simplify the sampling task.

4.2.1 *Matrix splitting*

We begin the review of MCMC samplers by detailing so-called *matrix splitting* (MS) approaches which build on the decomposition $\mathbf{Q} = \mathbf{M} - \mathbf{N}$

of the precision matrix. These methods embed one of the simplest and straightforward MCMC method to sample from a target Gaussian distribution, namely the component-wise Gibbs sampler (Geman and Geman, 1984).

Exact matrix splitting – Given the multivariate Gaussian distribution in (4.1), an attractive and simple option is to sequentially draw one component of $\boldsymbol{\theta}$ given the others. This is the well-known component-wise Gibbs sampler, see Algorithm 7 (Geman and Geman, 1984; Gelman et al., 2003; Rue and Held, 2005). The main advantage of Algorithm 7 is its simplicity and the low cost per *sweep* (i.e., internal iteration) of $\mathcal{O}(d^2)$ flops which is comparable with Cholesky applied to Toeplitz covariance matrices (Trench, 1964). More generally, one can also consider random sweeps over the d components of $\boldsymbol{\theta}$ or block-wise strategies which update simulteanously several components of $\boldsymbol{\theta}$. The analysis of these strategies and their respective convergence rates are detailed in the work by Roberts and Sahu (1997).

Algorithm 7: Component-wise Gibbs sampler

Input: Number T of iterations and initialization $\boldsymbol{\theta}^{(0)}$.

Set $t = 1$.
while $t \leq T$ **do**
  for $i \in [d]$ **do**
    Draw $z \sim \mathcal{N}(0, 1)$.
    Set $\theta_i^{(t)} = \frac{z}{\sqrt{Q_{ii}}} - \frac{1}{Q_{ii}} \left(\sum_{j>i} Q_{ij} \theta_j^{(t-1)} + \sum_{j<i} Q_{ij} \theta_j^{(t)} \right)$.
  end
  Set $t = t + 1$.
end

Output: $\boldsymbol{\mu} + \boldsymbol{\theta}^{(T)}$.

Adler (1981), Barone and Frigessi (1990), and Goodman and Sokal (1989) showed by rewriting Algorithm 7 using a matrix formulation that it actually stands for a stochastic sampling version of the Gauss-Seidel linear solver that relies on the decomposition $\mathbf{Q} = \mathbf{L} + \mathbf{D} + \mathbf{L}^T$ where $\mathbf{L}$ and $\mathbf{D}$ are the strictly lower triangular and diagonal parts of $\mathbf{Q}$, respectively. Indeed, each iteration solves the linear system

$$(\mathbf{L} + \mathbf{D})\boldsymbol{\theta}^{(t)} = \mathbf{D}^{1/2}\mathbf{z} - \mathbf{L}^T\boldsymbol{\theta}^{(t-1)} \tag{4.12}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. By setting $\mathbf{M} = \mathbf{L} + \mathbf{D}$ and $\mathbf{N} = -\mathbf{L}^T$ so that $\mathbf{Q} = \mathbf{M} - \mathbf{N}$, the updating rule (4.12) can be written as solving the usual Gauss-Seidel linear system:

$$\mathbf{M}\boldsymbol{\theta}^{(t)} = \tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)} \tag{4.13}$$

where $\mathbf{N} = -\mathbf{L}^T$ is strictly upper triangular and $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{D})$ is easy to sample.

Interestingly, (4.13) stands for a perturbed instance of MS schemes

which are a class of linear iterative solvers based on the splitting of $\mathbf{Q}$ into $\mathbf{Q} = \mathbf{M} - \mathbf{N}$ (Golub and Van Loan, 1989; Saad, 2003). Capitalizing on this one-to-one equivalence between Gibbs samplers and linear solvers, Fox and Parker (2017) extended Algorithm 7 to other Gibbs samplers based on different decompositions $\mathbf{Q} = \mathbf{M} - \mathbf{N}$. They are reported in Table 4.1 and yield Algorithm 8. Three important points can be noticed about

Algorithm 8: Gibbs sampler based on exact matrix splitting

Input: Number T of iterations, initialization $\boldsymbol{\theta}^{(0)}$ and splitting $\mathbf{Q} = \mathbf{M} - \mathbf{N}$.

1 Set $t = 1$.
2 **while** $t \leq T$ **do**
3 Draw $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{M}^T + \mathbf{N})$.
4 Solve $\mathbf{M}\boldsymbol{\theta}^{(t)} = \tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)}$ w.r.t. $\boldsymbol{\theta}^{(t)}$.
5 Set $t = t + 1$.
6 **end**

Output: $\boldsymbol{\mu} + \boldsymbol{\theta}^{(T)}$.

this algorithm. First, similarly to linear solvers, the sequence $\boldsymbol{\theta}^{(t)}$ built via Algorithm 8 is guaranteed to converge in distribution to π as $t \to \infty$ if $\rho\left(\mathbf{M}^{-1}\mathbf{N}\right) < 1$ where $\rho(\cdot)$ stands for the spectral radius of a matrix. Note that this is always the case for the component-wise Gibbs sampler although its convergence rate is generally low due to steps with very small variance. Second, the computational efficiency of Algorithm 8 is directly related to the complexity of solving the linear systems $\mathbf{M}\boldsymbol{\theta}^{(t)} = \tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)}$, similar to (4.13), and the difficulty of sampling $\tilde{\mathbf{z}}$ with covariance $\mathbf{M}^T + \mathbf{N}$. As pointed out by Fox and Parker (2017), the simpler $\mathbf{M}$, the denser $\mathbf{M}^T + \mathbf{N}$ and the more difficult the sampling of $\tilde{\mathbf{z}}$. In order to mitigate this trade-off, approximate MS approaches have been proposed recently (Barbos et al., 2017; Johnson, Saunderson, and Willsky, 2013). Finally, when the splitting is symmetric (both $\mathbf{M}$ and $\mathbf{N}$ are symmetric matrices), the rate of convergence of Algorithm 8 can be improved by using polynomial preconditioners, e.g., based on Chebyshev polynomials (Fox and Parker, 2017).

sampler	$\mathbf{M}$	$\mathbf{N}$	$\mathrm{var}(\tilde{\mathbf{z}}) = \mathbf{M}^T + \mathbf{N}$	convergence guarantee
Richardson	$\frac{1}{\omega}\mathbf{I}_d$	$\frac{1}{\omega}\mathbf{I}_d - \mathbf{Q}$	$\frac{2}{\omega}\mathbf{I}_d - \mathbf{Q}$	$0 < \omega < 2/\left\|\mathbf{Q}\right\|$
Jacobi	$\mathbf{D}$	$\mathbf{D} - \mathbf{Q}$	$2\mathbf{D} - \mathbf{Q}$	$\mathbf{Q}$ strictly diagonally dominant
Gauss-Seidel	$\mathbf{D} + \mathbf{L}$	$-\mathbf{L}^T$	$\mathbf{D}$	always
SOR	$\frac{1}{\omega}\mathbf{D} + \mathbf{L}$	$\frac{1-\omega}{\omega}\mathbf{D} - \mathbf{L}^T$	$\frac{2-\omega}{\omega}\mathbf{D}$	$0 < \omega < 2$

Table 4.1: Examples of MS schemes for $\mathbf{Q}$ which can be used in Algorithm 8. The matrices $\mathbf{D}$ and $\mathbf{L}$ denote the diagonal and strictly lower triangular parts of $\mathbf{Q}$, respectively. The vector $\tilde{\mathbf{z}}$ is the one appearing in step 3 of Algorithm 8 and ω is a positive scalar.

Approximate matrix splitting – Motivated by efficiency and parallel computations, Barbos et al. (2017) and Johnson, Saunderson, and Willsky (2013) proposed to relax exact MS and introduced two Gibbs samplers whose invariant distributions are approximations of π. First, in order to solve efficiently the linear system $\mathbf{M}\boldsymbol{\theta}^{(t)} = \tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)}$ involved in step 4 of Algorithm 8, these approximate approaches consider MS schemes with

diagonal matrices $\mathbf{M}$. For exact samplers, e.g., Richardson and Jacobi, we saw in the previous paragraph that such a convenient structure for $\mathbf{M}$ implies that the drawing of the Gaussian vector $\tilde{\mathbf{z}}$ becomes more demanding. To bypass this issue, approximate samplers draw Gaussian vectors $\tilde{\mathbf{z}}$ with simpler covariance matrices $\tilde{\mathbf{M}}$ instead of $\mathbf{M}^T + \mathbf{N}$. Again, attractive choices for $\tilde{\mathbf{M}}$ are diagonal matrices since the associated sampling task then boils down to Algorithm 4. This yields Algorithm 9 which is highly amenable to parallelization since both the covariance matrix $\tilde{\mathbf{M}}$ of $\tilde{\mathbf{z}}$ and the matrix $\mathbf{M}$ involved in the linear system to solve are diagonal.

Algorithm 9: Gibbs sampler based on approximate matrix splitting

Input: Number T of iterations, initialization $\boldsymbol{\theta}^{(0)}$ and splitting $\mathbf{Q} = \mathbf{M} - \mathbf{N}$.

1 Set $t = 1$.
2 **while** $t \leq T$ **do**
3 Draw $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}_d, \tilde{\mathbf{M}})$.
4 Solve $\mathbf{M}\boldsymbol{\theta}^{(t)} = \tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)}$ w.r.t. $\boldsymbol{\theta}^{(t)}$.
5 Set $t = t + 1$.
6 **end**

Output: $\boldsymbol{\mu} + \boldsymbol{\theta}^{(T)}$.

Table 4.2 gathers the respective expressions of $\mathbf{M}$, $\mathbf{N}$ and $\tilde{\mathbf{M}}$ for the two approaches introduced by Johnson, Saunderson, and Willsky (2013) and by Barbos et al. (2017). These two approaches define a Markov chain whose invariant distribution is a Gaussian with the correct mean $\boldsymbol{\mu}$ but with precision matrix $\widetilde{\mathbf{Q}}$, where

$$\widetilde{\mathbf{Q}} = \begin{cases} \mathbf{Q}\left(\mathbf{I}_d - \mathbf{D}^{-1}(\mathbf{L} + \mathbf{L}^T)\right) & \text{for the Hogwild sampler} \\ \mathbf{Q}\left(\mathbf{I}_d - \frac{1}{2}(\mathbf{D} + 2\omega^{-1}\mathbf{I}_d)^{-1}\mathbf{Q}\right) & \text{for clone MCMC.} \end{cases} \tag{4.14}$$

Contrary to the Hogwild sampler, clone MCMC is able to sample exactly from π in the asymptotic scenario $\omega \to 0$ since in this case $\widetilde{\mathbf{Q}} \to \mathbf{Q}$.

Interestingly, these state-of-the-art MCMC schemes are two special instances of the unifying framework proposed in Section 4.3 as well as the AXDA one introduced in Chapter 1. These connections will be further highlighted in Section 4.3.

sampler	$\mathbf{M}$	$\mathbf{N}$	$\mathrm{cov}(\tilde{\mathbf{z}}) = \tilde{\mathbf{M}}$
Hogwild with blocks of size 1	$\mathbf{D}$	$-\mathbf{L} - \mathbf{L}^T$	$\mathbf{D}$
Clone MCMC	$\mathbf{D} + 2\omega\mathbf{I}_d$	$2\omega\mathbf{I}_d - \mathbf{L} - \mathbf{L}^T$	$2\left(\mathbf{D} + 2\omega\mathbf{I}_d\right)$

Table 4.2: MS schemes for $\mathbf{Q}$ which can be used in Algorithm 9. The matrices $\mathbf{D}$ and $\mathbf{L}$ denote the diagonal and strictly lower triangular parts of $\mathbf{Q}$, respectively. The vector $\tilde{\mathbf{z}}$ is the one appearing in step 3 of Algorithm 9 and $\omega > 0$ is a tuning parameter controlling the bias of those methods. Sufficient conditions to guarantee $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$ are given in (Johnson, Saunderson, and Willsky, 2013; Barbos et al., 2017).

4.2.2 *Data augmentation*

Since the precision matrix $\mathbf{Q}$ has been assumed to be arbitrary, the MS schemes $\mathbf{Q} = \mathbf{M} - \mathbf{N}$ in Table 4.1 were not motivated by its structure but rather by the computational efficiency of the associated samplers. Hence,

inspired by efficient linear solvers, relevant choices for $\mathbf{M}$ and $\mathbf{N}$ given in Table 4.1 and Table 4.2 have been considered. Another line of research explores schemes specifically dedicated to precision matrices $\mathbf{Q}$ of the form

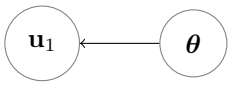
$$\mathbf{Q} = \mathbf{Q}_1 + \mathbf{Q}_2, \tag{4.15}$$

where, contrary to the MS schemes discussed in the previous section, the two matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are not chosen by the user but directly result from the statistical model under consideration. In particular, such situations arise when deriving hierarchical Bayesian models (see, e.g., the works by Rue and Held (2005), Idier (2008), and Orieux, Giovannelli, and Rodet (2010)). By capitalizing on possible specific structures of $\{\mathbf{Q}_i\}_{i\in[2]}$, it may be desirable to separate $\mathbf{Q}_1$ and $\mathbf{Q}_2$ in two different hopefully simpler steps of a Gibbs sampler. To this purpose, this section discusses *data augmentation* (DA) approaches which introduce one (or several) auxiliary variable $\mathbf{u} \in \mathbb{R}^k$ such that the joint distribution of the couple $(\boldsymbol{\theta}, \mathbf{u})$ yields simple conditional distributions thus sampling steps within a Gibbs sampler (Barbos et al., 2017; Marnissi et al., 2018; Vono, Dobigeon, and Chainais, 2019a; Marnissi et al., 2019). Then a straightforward marginalization of the auxiliary variable $\mathbf{u}$ permits to retrieve the distribution π, either exactly or in an asymptotic regime depending on the nature of the DA scheme.

Exact data augmentation – This paragraph reviews some exact DA approaches to obtain samples from π. The term *exact* means here that the joint distribution $\pi(\boldsymbol{\theta}, \mathbf{u})$ satisfies almost surely

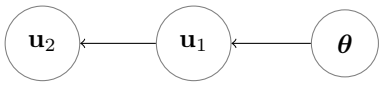
$$\int_{\mathbb{R}^k} \pi(\boldsymbol{\theta}, \mathbf{u}) \mathrm{d}\mathbf{u} = \pi(\boldsymbol{\theta}) \tag{4.16}$$

and yields proper marginal distributions $\pi(\boldsymbol{\theta})$ and $\pi(\mathbf{u})$. Figure 4.4 describes the directed acyclic graphs (DAG) associated with two hierarchical models proposed by Marnissi et al. (2018) and Marnissi et al. (2019) to decouple $\mathbf{Q}_1$ from $\mathbf{Q}_2$ by involving auxiliary variables. In the following,



$$\mathbf{u}_1|\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\theta}, \mathbf{P}^{-1}\right),\ \mathbf{P} = \omega^{-1}\mathbf{I}_d - \mathbf{Q}_1$$

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{Q}^{-1}\right)$$



$$\mathbf{u}_2|\mathbf{u}_1 \sim \mathcal{N}\left(\mathbf{G}_1\mathbf{u}_1, \mathbf{\Lambda}_1^{-1}\right)$$

$$\mathbf{u}_1|\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\theta}, \mathbf{P}^{-1}\right)$$

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{Q}^{-1}\right)$$

Figure 4.4: Hierarchical models proposed by Marnissi et al., 2018; Marnissi et al., 2019 where ω is such that $0 < \omega < \|\mathbf{Q}_1\|^{-1}$.

we detail the motivations behind these two data augmentation schemes. Among the two matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ involved in the composite precision matrix $\mathbf{Q}$, without loss of generality, we assume that $\mathbf{Q}_2$ presents a particular and simpler structure (e.g., diagonal or circulant) than $\mathbf{Q}_1$. We want now to benefit from this structure by leveraging on the efficient sampling schemes previously discussed in Section 4.1.2 and well suited to handle a Gaussian distribution with a precision matrix only involving $\mathbf{Q}_2$. This is the aim of the first data augmentation model called EDA which introduces

the joint distribution

$$\pi(\boldsymbol{\theta}, \mathbf{u}_1) \propto \exp\left(-\frac{1}{2}\left[(\boldsymbol{\theta}-\boldsymbol{\mu})^T\mathbf{Q}(\boldsymbol{\theta}-\boldsymbol{\mu}) + (\mathbf{u}_1-\boldsymbol{\theta})^T\mathbf{P}(\mathbf{u}_1-\boldsymbol{\theta})\right]\right) \tag{4.17}$$

with $\mathbf{P} = \omega^{-1}\mathbf{I}_d - \mathbf{Q}_1$ and $0 < \omega < \|\mathbf{Q}_1\|^{-1}$, where $\|\cdot\|$ is the spectral norm of a matrix. The resulting Gibbs sampler (see Algorithm 10) relies on two conditional Gaussian sampling steps whose associated conditional distributions are detailed in Table 4.3.

Algorithm 10: Gibbs sampler based on exact data augmentation

Input: Number T of iterations and initialization $\boldsymbol{\theta}^{(0)}$, $\mathbf{u}_1^{(0)}$.

Set $t = 1$.
while $t \leq T$ **do**
  Draw $\mathbf{u}_2^{(t)} \sim \mathcal{N}(\mu_{\mathbf{u}_2}, \mathbf{Q}_{\mathbf{u}_2}^{-1})$. `// Only if GEDA is considered.`
  Draw $\mathbf{u}_1^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}_1}, \mathbf{Q}_{\mathbf{u}_1}^{-1})$.
  Draw $\boldsymbol{\theta}^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}}^{-1})$.
  Set $t = t + 1$.
end

Output: $\boldsymbol{\mu} + \boldsymbol{\theta}^{(T)}$.

sampler	$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}}^{-1})$	$\mathbf{u}_1 \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}_1}, \mathbf{Q}_{\mathbf{u}_1}^{-1})$	$\mathbf{u}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}_2}, \mathbf{Q}_{\mathbf{u}_2}^{-1})$
EDA	$\mathbf{Q}_{\boldsymbol{\theta}} = \omega^{-1}\mathbf{I}_d + \mathbf{Q}_2$ $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{Q}_{\boldsymbol{\theta}}^{-1}(\mathbf{P}\mathbf{u}_1 + \mathbf{Q}\boldsymbol{\mu})$	$\mathbf{Q}_{\mathbf{u}_1} = \mathbf{P}$ $\boldsymbol{\mu}_{\mathbf{u}_1} = \boldsymbol{\theta}$	- -
GEDA	$\mathbf{Q}_{\boldsymbol{\theta}} = \omega^{-1}\mathbf{I}_d + \mathbf{Q}_2$ $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{Q}_{\boldsymbol{\theta}}^{-1}(\mathbf{P}\mathbf{u}_1 + \mathbf{Q}\boldsymbol{\mu})$	$\mathbf{Q}_{\mathbf{u}_1} = \omega^{-1}\mathbf{I}_d$ $\boldsymbol{\mu}_{\mathbf{u}_1} = \boldsymbol{\theta} - \omega(\mathbf{Q}_1\boldsymbol{\theta} - \mathbf{G}_1^T\boldsymbol{\Lambda}_1^{-1}\mathbf{u}_2)$	$\mathbf{Q}_{\mathbf{u}_2} = \boldsymbol{\Lambda}_1$ $\boldsymbol{\mu}_{\mathbf{u}_2} = \mathbf{G}_1\mathbf{u}_1$

Table 4.3: Conditional probability distributions for exact data augmentation schemes. The parameter ω is such that $0 < \omega < \|\mathbf{Q}_1\|^{-1}$. For simplicity, the conditioning is notationally omitted.

This scheme has the great advantage of decoupling the two precision matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ since they are not simultaneously involved in any of the two steps. In particular, introducing the auxiliary variable $\mathbf{u}_1$ permits to remove the dependence in $\mathbf{Q}_1$ when defining the precision matrix of the conditional distribution of $\boldsymbol{\theta}$. While efficient sampling from this conditional is now possible, we have to ensure that sampling the auxiliary variable $\mathbf{u}_1$ can be achieved with a reasonable computational cost. Again if $\mathbf{Q}_1$ presents a nice structure, the specific approaches reviewed in Section 4.1.2 can be employed. If this is not the case, Marnissi et al. (2018) and Marnissi et al. (2019) proposed a generalization of EDA, called GEDA, to simplify the whole Gibbs sampling procedure when $\mathbf{Q}$ arises from a hierarchical Bayesian model. In such models, $\mathbf{Q}_1$ (and a fortiori $\mathbf{Q}_2$) naturally admits an explicit decomposition which writes $\mathbf{Q}_1 = \mathbf{G}_1^T\boldsymbol{\Lambda}_1\mathbf{G}_1$, where $\boldsymbol{\Lambda}_1$ is a positive definite (and very often diagonal) matrix. By building on this explicit decomposition, GEDA introduces an additional auxiliary variable

$\mathbf{u}_2$ such that

$$\pi(\boldsymbol{\theta}, \mathbf{u}_1, \mathbf{u}_2) \propto \exp\left(-\frac{1}{2}\left[(\boldsymbol{\theta}-\boldsymbol{\mu})^T\mathbf{Q}(\boldsymbol{\theta}-\boldsymbol{\mu}) + (\mathbf{u}_1-\boldsymbol{\theta})^T\mathbf{P}(\mathbf{u}_1-\boldsymbol{\theta})\right]\right) \times \exp\left(-\frac{1}{2}(\mathbf{u}_2-\mathbf{G}_1\mathbf{u}_1)^T\boldsymbol{\Lambda}_1(\mathbf{u}_2-\mathbf{G}_1\mathbf{u}_1)\right). \quad (4.18)$$

This joint distribution yields conditional Gaussian distributions with diagonal covariance matrices for both $\mathbf{u}_1$ and $\mathbf{u}_2$ that can be sampled efficiently, see Table 4.3.

Approximate data augmentation – The AXDA framework was introduced in Chapter 1 to handle any target distributions and therefore applies to the Gaussian case as well, as already illustrated in Section 2.2.1 of Chapter 2. In what follows, we implement this framework to this special case. An auxiliary variable $\mathbf{u} \in \mathbb{R}^d$ can be introduced such that the joint pdf of $(\boldsymbol{\theta}, \mathbf{u})$ writes

$$\pi(\boldsymbol{\theta}, \mathbf{u}) \propto \exp\left(-\frac{1}{2}\left[(\boldsymbol{\theta}-\boldsymbol{\mu})^T\mathbf{Q}_2(\boldsymbol{\theta}-\boldsymbol{\mu}) + (\mathbf{u}-\boldsymbol{\mu})^T\mathbf{Q}_1(\mathbf{u}-\boldsymbol{\mu}) + \frac{\|\mathbf{u}-\boldsymbol{\theta}\|^2}{\omega}\right]\right) \quad (4.19)$$

where $\omega > 0$. The main idea behind (4.19) is to replicate the variable of interest $\boldsymbol{\theta}$ in order to sample two different random variables $\mathbf{u}$ and $\boldsymbol{\theta}$ with covariances involving separately $\mathbf{Q}_1$ and $\mathbf{Q}_2$. The marginal $\pi(\boldsymbol{\theta})$ under the joint $\pi(\boldsymbol{\theta}, \mathbf{u})$ in (4.19) is a Gaussian with the correct mean $\boldsymbol{\mu}$ but with an approximate precision matrix $\widetilde{\mathbf{Q}}$ which admits the closed-form expression

$$\widetilde{\mathbf{Q}} = \mathbf{Q}_2 + \left(\mathbf{Q}_1^{-1} + \omega\mathbf{I}_d\right)^{-1}. \quad (4.20)$$

For $\omega > 0$, approximate samples from π can be generated by SGS that sequentially draws from the conditional distributions

$$\mathbf{u}|\boldsymbol{\theta} \sim \mathcal{N}\left((\omega^{-1}\mathbf{I}_d + \mathbf{Q}_1)^{-1}(\omega^{-1}\boldsymbol{\theta} + \mathbf{Q}_1\boldsymbol{\mu}), (\omega^{-1}\mathbf{I}_d + \mathbf{Q}_1)^{-1}\right) \quad (4.21)$$

$$\boldsymbol{\theta}|\mathbf{u} \sim \mathcal{N}\left((\omega^{-1}\mathbf{I}_d + \mathbf{Q}_2)^{-1}(\omega^{-1}\mathbf{u} + \mathbf{Q}_2\boldsymbol{\mu}), (\omega^{-1}\mathbf{I}_d + \mathbf{Q}_2)^{-1}\right). \quad (4.22)$$

Again, this approach has the great advantage of decoupling the two precision matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ defining $\mathbf{Q}$ since they are not simultaneously involved in any of the two steps of the Gibbs sampler. Marnissi et al. (2019) showed that exact DA schemes (i.e., EDA and GEDA) generally outperform AXDA as far as Gaussian sampling is concerned. This was expected since the AXDA framework proposed is not specifically designed for Gaussian targets but for a wide family of distributions.

4.3 A unifying approach

This section proposes to unify and extend most of the MCMC approaches detailed in Section 4.2 by building upon a general Gaussian simulation framework which can be interestingly seen as a stochastic sampling coun-

terpart of the celebrated proximal point algorithm (PPA) (Rockafellar, 1976), see Section 4.4. This viewpoint will shed new light on the connections between existing simulation-based algorithms, and particularly between Gibbs samplers.

4.3.1 *A unifying proposal distribution*

In Section 4.1.3, we highlighted that the main difficulty related to the considered Gaussian sampling problem is to handle the high-dimensional precision matrix $\mathbf{Q}$. In the sequel, we propose to bypass this issue by relying on a class of surrogate probability distributions (e.g., conditional or approximate distributions) to make Gaussian sampling easier. We model this idea by considering a general probability distribution with density κ such that

$$\kappa(\boldsymbol{\theta}, \mathbf{u}) \propto \pi(\boldsymbol{\theta}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\mathbf{u})^T\mathbf{P}(\boldsymbol{\theta}-\mathbf{u})\right) \tag{4.23}$$

where $\pi(\boldsymbol{\theta})$ is the target Gaussian distribution, $\mathbf{u} \in \mathbb{R}^d$ stands for an additional (auxiliary) variable and $\mathbf{P} \in \mathbb{R}^{d\times d}$ is a symmetric matrix acting as a preconditioner such that κ defines a proper density on an appropriate state space. More precisely, in the following and with a slight abuse of notations to ease the presentation, depending on the definition of the variable $\mathbf{u}$, the probability density κ in (4.23) shall refer to either the conditional probability density $\pi(\boldsymbol{\theta}|\mathbf{u})$ or the joint pdf $\pi(\boldsymbol{\theta}, \mathbf{u})$.

This appropriate state space will be made explicit in the following sections.

4.3.2 *From exact data augmentation to exact matrix splitting*

We assume here that the variable $\mathbf{u}$ refers to an auxiliary variable such that the joint distribution of the couple $(\boldsymbol{\theta}, \mathbf{u})$ has a density given by $\pi(\boldsymbol{\theta}, \mathbf{u}) \coloneqq \kappa(\boldsymbol{\theta}, \mathbf{u})$. In addition, we restrict here $\mathbf{P}$ to be positive definite. It follows that

$$\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}, \mathbf{u})\mathrm{d}\mathbf{u} = Z^{-1}\pi(\boldsymbol{\theta}) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\mathbf{u})^T\mathbf{P}(\boldsymbol{\theta}-\mathbf{u})\right) \mathrm{d}\mathbf{u} = \pi(\boldsymbol{\theta}) \tag{4.24}$$

holds almost surely with $Z = \det(\mathbf{P})^{-1/2}(2\pi)^{d/2} < \infty$. Hence, the joint density (4.23) yields an exact DA scheme whatever the choice of the positive definite matrix $\mathbf{P}$. We will show that the exact DA approaches schemed by Algorithm 10 precisely fit the proposed generic framework with a specific choice for the preconditioning matrix $\mathbf{P}$. We will then extend this class of exact DA approaches and show a one-to-one equivalence between Gibbs samplers based on exact MS and those based on exact DA.

Note that $Z < \infty$ because we assumed that $\mathbf{P}$ is positive definite which yields $\det(\mathbf{P}) > 0$.

To this purpose, we start by making the change of variable $\mathbf{v} = \mathbf{P}\mathbf{u}$. Combined with the joint probability density (4.23), it yields the two following conditional probability densities:

$$\pi(\mathbf{v}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{P}\boldsymbol{\theta}, \mathbf{P}\right), \tag{4.25}$$

$$\pi(\boldsymbol{\theta}|\mathbf{v}) = \mathcal{N}\left((\mathbf{Q}+\mathbf{P})^{-1}(\mathbf{v}+\mathbf{Q}\boldsymbol{\mu}), (\mathbf{Q}+\mathbf{P})^{-1}\right). \tag{4.26}$$

By re-writing the Gibbs sampling steps associated to these two conditionals as an auto-regressive process of order 1 w.r.t. $\boldsymbol{\theta}$ (Box and Jenkins, 1994), it follows that an equivalent sampling strategy writes

$$\tilde{\mathbf{z}} \sim \mathcal{N}\left(\mathbf{Q}\boldsymbol{\mu}, 2\mathbf{P} + \mathbf{Q}\right), \tag{4.27}$$

$$\boldsymbol{\theta}^{(t)} = \left(\mathbf{Q} + \mathbf{P}\right)^{-1}\left(\tilde{\mathbf{z}} + \mathbf{P}\boldsymbol{\theta}^{(t-1)}\right). \tag{4.28}$$

Defining $\mathbf{M} = \mathbf{Q} + \mathbf{P}$ and $\mathbf{N} = \mathbf{P}$, or equivalently $\mathbf{Q} = \mathbf{M} - \mathbf{N}$, it yields

$$\tilde{\mathbf{z}} \sim \mathcal{N}\left(\mathbf{Q}\boldsymbol{\mu}, \mathbf{M}^T + \mathbf{N}\right), \tag{4.29}$$

$$\boldsymbol{\theta}^{(t)} = \mathbf{M}^{-1}\left(\tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)}\right) \tag{4.30}$$

which boils down to a Gibbs sampler based on exact MS (see Section 4.3.3). Such samplers have been proposed and studied by Fox and Parker (2017). The mean vector $\boldsymbol{\mu}$ appears here in (4.29) instead of being added at the end of the sampling similarly to the approaches reviewed before. Note however that it is equivalent to consider a zero-mean Gaussian vector $\tilde{\mathbf{z}}$ within the Gibbs sampling procedure and to add $\boldsymbol{\mu}$ at the end.

To illustrate the interest of this rewriting when considering the case of two matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ that cannot be efficiently handled in the same basis, Table 4.4 presents two possible choices of $\mathbf{P}$ which relate two MS strategies with their DA counterparts. Firstly, one particular choice of $\mathbf{P}$ (row 1 of Table 4.4) directly shows that the Richardson's MS sampler proposed by Fox and Parker (2017) can be rewritten as the EDA sampler. More precisely, the auto-regressive process of order 1 w.r.t. $\boldsymbol{\theta}$ defined by EDA yields a variant of the Richardson's sampler. This finding relates two different approaches proposed by authors from distinct communities (numerical linear algebra and signal processing). Secondly, the proposed unifying framework also permits to go beyond existing approaches by proposing a novel exact DA approach via a specific choice for the precision matrix $\mathbf{P}$ driven by an existing MS method. Indeed, following the same rewritting trick with another particular choice of $\mathbf{P}$ (row 2 of Table 4.4), an exact DA scheme can be easily derived from the Jacobi's MS approach. Up to our knowledge, this novel DA method, referred to as EDAJ in the table, has not been documented in the existing literature.

Finally, this table reports two particular choices of $\mathbf{P}$ which lead to revisit existing MS and/or DA methods. It is worth noting that other relevant choices may be possible, which would allow to derive new exact DA and MS methods or to draw further analogies between existing approaches. Note also that Table 4.4 shows the main benefit of an exact DA scheme over its MS counterpart thanks to the decoupling between $\mathbf{Q}_1$ and $\mathbf{Q}_2$ in two separate simulation steps. This feature can be directly observed by comparing the two first columns of Table 4.4 with the third one.

$\mathbf{P} = \text{cov}(\mathbf{v}\|\boldsymbol{\theta})$	$(\mathbf{Q}+\mathbf{P})^{-1} = \text{cov}(\boldsymbol{\theta}\|\mathbf{v})$	$\mathbf{M}^T + \mathbf{N} = \text{cov}(\tilde{\mathbf{z}})$	DA sampler	MS sampler
$\frac{\mathbf{I}_d}{\omega} - \mathbf{Q}_1$	$\left(\frac{\mathbf{I}_d}{\omega} + \mathbf{Q}_2\right)^{-1}$	$\frac{2\mathbf{I}_d}{\omega} + \mathbf{Q}_2 - \mathbf{Q}_1$	EDA	Richardson
$\frac{\mathbf{D}}{\omega} - \mathbf{Q}_1$	$\left(\frac{\mathbf{D}}{\omega} + \mathbf{Q}_2\right)^{-1}$	$\frac{2\mathbf{D}}{\omega} + \mathbf{Q}_2 - \mathbf{Q}_1$	EDAJ	Jacobi

Table 4.4: Equivalence relations between exact DA and exact MS approaches. The matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are such that $\mathbf{Q} = \mathbf{Q}_1 + \mathbf{Q}_2$. The matrices $\mathbf{D}$ and $\mathbf{L}$ denote the diagonal and strictly lower triangular parts of $\mathbf{Q}_1$, respectively, and $\omega > 0$ is a positive scalar ensuring the positive definiteness of $\mathbf{P}$.

4.3.3 *From approximate matrix splitting to approximate data augmentation*

We now build on the proposed unifying proposal (4.23) to extend the class of samplers based on approximate matrix splitting and reviewed in Section 4.2.1. More precisely, let define $\mathbf{u} = \boldsymbol{\theta}^{(t-1)}$ to be the current iterate within an MCMC algorithm and κ to be

$$\kappa(\boldsymbol{\theta}, \mathbf{u}) = \pi\left(\boldsymbol{\theta}|\mathbf{u} = \boldsymbol{\theta}^{(t-1)}\right) \propto \pi(\boldsymbol{\theta}) \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)}\right)^T \mathbf{P}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)}\right)\right). \tag{4.31}$$

Readers familiar with MCMC algorithms will recognize in (4.31) a proposal distribution that can be used within Metropolis-Hastings schemes (Robert and Casella, 2004). However, unlike the usual random-walk algorithm which considers the Gaussian proposal $\mathcal{N}(\boldsymbol{\theta}^{(t-1)}, \lambda\mathbf{I}_d)$ with $\lambda > 0$, the originality of (4.31) is to define the proposal by combining the Gaussian target π with a term that is equal to a Gaussian kernel when $\mathbf{P}$ is positive definite. If we always accept the proposed sample obtained from (4.31) without any correction, that is $\boldsymbol{\theta}^{(t)} = \widetilde{\boldsymbol{\theta}} \sim \pi(\widetilde{\boldsymbol{\theta}}|\mathbf{u} = \boldsymbol{\theta}^{(t-1)})$, this directly implies that the associated Markov chain converges in distribution towards a Gaussian random variable with distribution $\tilde{\pi}$ with the correct mean $\boldsymbol{\mu}$ but with precision matrix

$$\widetilde{\mathbf{Q}} = \mathbf{Q}\left(\mathbf{I}_d + (\mathbf{P}+\mathbf{Q})^{-1}\mathbf{P}\right). \tag{4.32}$$

This algorithm is detailed in Algorithm 11. Note again that one can obtain samples from the initial target distribution π by replacing step 4 with an acceptance/rejection step, see the textbook by Robert and Casella (2004) for details.

Algorithm 11: Gibbs sampler based on (4.31).

Input: Number T of iterations and initialization $\boldsymbol{\theta}^{(0)}$.

1 Set $t = 1$.
2 **while** $t \leq T$ **do**
3 Draw $\widetilde{\boldsymbol{\theta}} \sim \pi\left(\widetilde{\boldsymbol{\theta}}|\mathbf{u} = \boldsymbol{\theta}^{(t-1)}\right)$ // see (4.31).
4 Set $\boldsymbol{\theta}^{(t)} = \widetilde{\boldsymbol{\theta}}$.
5 Set $t = t + 1$.
6 **end**

Output: $\boldsymbol{\mu} + \boldsymbol{\theta}^{(T)}$.

The instance (4.31) of (4.23) paves the way to an extended class of samplers based on approximate MS. More precisely, since $\pi = \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$, the draw of a proposed sample $\widetilde{\boldsymbol{\theta}}$ from (4.31) can be replaced by the

following two-step sampling procedure:

$$\tilde{\mathbf{z}} \sim \mathcal{N}\left(\mathbf{Q}\boldsymbol{\mu}, \mathbf{P}+\mathbf{Q}\right), \tag{4.33}$$

$$\boldsymbol{\theta}^{(t)} = \left(\mathbf{Q}+\mathbf{P}\right)^{-1}\left(\tilde{\mathbf{z}} + \mathbf{P}\boldsymbol{\theta}^{(t-1)}\right). \tag{4.34}$$

The matrix splitting form with $\mathbf{M} = \mathbf{Q}+\mathbf{P}$, $\mathbf{N} = \mathbf{P}$ writes

$$\tilde{\mathbf{z}} \sim \mathcal{N}\left(\mathbf{Q}\boldsymbol{\mu}, \mathbf{M}\right), \tag{4.35}$$

$$\boldsymbol{\theta}^{(t)} = \mathbf{M}^{-1}\left(\tilde{\mathbf{z}} + \mathbf{N}\boldsymbol{\theta}^{(t-1)}\right). \tag{4.36}$$

This recursion defines an extended class of approximate MS-based samplers and encompasses the so-called Hogwild sampler proposed by Johnson, Saunderson, and Willsky (2013) by taking $\mathbf{P} = -\mathbf{L} - \mathbf{L}^T$, where $\mathbf{L}$ stands for the strictly lower triangular part of $\mathbf{Q}$. In addition to the existing Hogwild approach, Table 4.5 lists two other and new approximate MS approaches resulting from specific choices of the preconditioning matrix $\mathbf{P}$. They are coined *approximate* Richardson and Jacobi samplers since the expressions for $\mathbf{M}$ and $\mathbf{N}$ are very similar to the ones associated to their exact counterparts, see Fox and Parker (2017). For those two samplers, note that the approximate precision matrix $\widetilde{\mathbf{Q}}$ tends towards $2\mathbf{Q}$ in the asymptotic regime $\omega \to 0$. Indeed, for the approximate Jacobi sampler, we have for instance

$$\widetilde{\mathbf{Q}} = \mathbf{Q}\left(\mathbf{I}_d + \omega\left(\frac{\mathbf{I}_d}{\omega} - \mathbf{Q}\right)\right) \tag{4.37}$$

$$= \mathbf{Q}\left(2\mathbf{I}_d - \omega\mathbf{Q}\right) \tag{4.38}$$

$$\underset{\omega\to 0}{\to} 2\mathbf{Q}. \tag{4.39}$$

In order to retrieve the original precision matrix $\mathbf{Q}$ when $\omega \to 0$, Barbos et al. (2017) proposed an approximate data augmentation strategy which can be related to the AXDA framework introduced in Chapter 1.

$\frac{1}{2}\mathbf{M} = \mathrm{cov}(\mathbf{v}\|\boldsymbol{\theta})$	$\frac{1}{2}\mathbf{M}^{-1} = \mathrm{cov}(\boldsymbol{\theta}\|\mathbf{v})$	$\mathbf{M} = \mathrm{cov}(\tilde{\mathbf{z}})$	MS sampler	DA sampler
$\frac{1}{2}\mathbf{D}$	$\frac{1}{2}\mathbf{D}^{-1}$	$\mathbf{D}$	Hogwild with blocks of size 1	ADAH
$\frac{\mathbf{I}_d}{2\omega}$	$\frac{\omega\mathbf{I}_d}{2}$	$\frac{\mathbf{I}_d}{\omega}$	approx. Richardson	ADAR
$\frac{\mathbf{D}}{2\omega}$	$\frac{\omega\mathbf{D}^{-1}}{2}$	$\frac{\mathbf{D}}{\omega}$	approx. Jacobi	ADAJ

Table 4.5: Extended class of Gibbs samplers based on approximate MS with $\mathbf{Q} = \mathbf{M} - \mathbf{N}$ with $\mathbf{N} = \mathbf{P}$ and approximate DA. The matrices $\mathbf{D}$ and $\mathbf{L}$ denote the diagonal and strictly lower triangular parts of $\mathbf{Q}$, respectively. ω is a positive scalar.

In the following, we will show that approximate MS approaches admit approximate DA counterparts, which are highly amenable to distributed and parallel computations. The recursion (4.36) can be equivalently written as

$$\widetilde{\boldsymbol{\theta}} = \left(\mathbf{Q}+\mathbf{P}\right)^{-1}\left(\mathbf{Q}\boldsymbol{\mu} + \mathbf{P}\boldsymbol{\theta}^{(t-1)} + \mathbf{z}_1\right) + \mathbf{z}_2, \tag{4.40}$$

where $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{2}(\mathbf{P}+\mathbf{Q}))$ and $\mathbf{z}_2 \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{2}(\mathbf{P}+\mathbf{Q})^{-1})$. By introducing an auxiliary variable $\mathbf{v}$ defined by $\mathbf{v} = \mathbf{P}\boldsymbol{\theta}^{(t-1)} + \mathbf{z}_1$, the resulting

two-step Gibbs sampling relies on the conditional distributions

$$\mathbf{v}|\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{P}\boldsymbol{\theta}, \frac{1}{2}(\mathbf{P}+\mathbf{Q})\right), \tag{4.41}$$

$$\boldsymbol{\theta}|\mathbf{v} \sim \mathcal{N}\left((\mathbf{Q}+\mathbf{P})^{-1}(\mathbf{v}+\mathbf{Q}\boldsymbol{\mu}), \frac{1}{2}(\mathbf{P}+\mathbf{Q})^{-1}\right) \tag{4.42}$$

and targets the joint density $\pi(\boldsymbol{\theta}, \mathbf{v})$. Interestingly, the sampling difficulty associated to each conditional sampling step is the same and only driven by the structure of the matrix $\mathbf{M} = \mathbf{P} + \mathbf{Q}$. In particular, this matrix becomes diagonal for three specific choices listed in Table 4.5. These choices lead to three sampling schemes, referred to as ADAH, ADAR and ADAJ, which are the DA counterparts of the approximate MS samplers discussed above. These DA schemes have the great advantage of leading to Gibbs samplers suited for parallel computations, hence simplifying the sampling procedure. Contrary to exact approaches detailed in Section 4.3.2, note that these DA schemes naturally emerge here without assuming any explicit decomposition $\mathbf{Q} = \mathbf{Q}_1 + \mathbf{Q}_2$ or including an additional auxiliary variable.

4.4 Gibbs samplers as stochastic sampling counterparts of the PPA

This section aims at drawing new connections between sampling and optimization approaches. More precisely, we will show that approximate Gibbs samplers based on the proposal (4.31) can be interestingly seen as stochastic sampling counterparts of the celebrated proximal point algorithm (PPA) in optimization (Rockafellar, 1976). We assume here that $\mathbf{P}$ is positive semi-definite and define the *weighted* norm w.r.t. $\mathbf{P}$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$ by

$$\|\boldsymbol{\theta}\|_{\mathbf{P}} \coloneqq \sqrt{\boldsymbol{\theta}^T \mathbf{P} \boldsymbol{\theta}}. \tag{4.43}$$

4.4.1 *The proximal point algorithm*

Let $\mathcal{H}$ a real Hilbert space. The PPA is an important and widely used method to find zeros of a maximal monotone operator $\mathsf{K} : \mathcal{H} \to 2^{\mathcal{H}}$, that is to solve problems of the form

The notation $2^{\mathcal{H}}$ stands for the family of all subsets of $\mathcal{H}$, see Bauschke and Combettes (2013) for more details. This implies that K stands for a set-valued operator.

$$\text{Find } \boldsymbol{\theta}^\star \in \mathcal{H} \text{ such that } \mathbf{0}_d \in \mathsf{K}(\boldsymbol{\theta}^\star), \tag{4.44}$$

where $\mathcal{H}$ is a real Hilbert space. For simplicity, we will take here $\mathcal{H} = \mathbb{R}^d$ equipped with the usual Euclidean norm and focus on the particular case $\mathsf{K} = \partial f$ where f is a lower semicontinuous (l.s.c.), proper, coercive and convex function and ∂ denotes the subdifferential operator. In this case, the PPA is equivalent to the proximal minimization algorithm (Martinet, 1970, 1972) which aims at solving the minimization problem

$$\text{Find } \boldsymbol{\theta}^\star \in \mathbb{R}^d \text{ such that } \boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} f(\boldsymbol{\theta}), \tag{4.45}$$

by generating a sequence $\{\boldsymbol{\theta}^{(t)}\}_{t\in\mathbb{N}}$ which solves successive approximations of the minimization problem (4.45), i.e., for $t \in \mathbb{N}^*$ and $\lambda > 0$,

$$\begin{aligned}\boldsymbol{\theta}^{(t)} &= \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} f(\boldsymbol{\theta}) + \frac{1}{2\lambda}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)}\right\|^2. \\ &:= \operatorname{prox}_{\lambda f}(\boldsymbol{\theta}^{(t-1)}).\end{aligned} \tag{4.46}$$

This algorithm is called the proximal point algorithm in reference to the work by Moreau (1965). When $\mathcal{H}$ is equipped with $\langle\cdot,\cdot\rangle_{\mathbf{P}}$, the PPA is detailed in Algorithm 12. This algorithm can be dated back at least to

Algorithm 12: Proximal point algorithm (PPA) with $\langle\cdot,\cdot\rangle_{\mathbf{P}}$

Input: Choose an initial value $\boldsymbol{\theta}^{(0)}$, a positive semi-definite matrix $\mathbf{P}$ and a maximal number of iterations T.

1 Set $t = 1$.
2 **while** $t \leq T$ **do**
3 $\quad \boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\ f(\boldsymbol{\theta}) + \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)}\right\|_{\mathbf{P}}^2$.
4 $\quad$ Set $t = t + 1$.
5 **end**

Output: $\boldsymbol{\theta}^{(T)}$.

Bellman, Kalaba, and Lockett (1966) who considered successive approximations of an initial quadratic problem to solve efficiently ill-conditionned linear systems. Note that instead of directly minimizing the objective function f, Algorithm 12 successively adds a quadratic penalty term depending on the previous iterate $\boldsymbol{\theta}^{(t-1)}$ and then solves an approximation of the initial optimization problem at each iteration. This idea of successive approximations is exactly the deterministic counterpart of (4.31) which proposes a new sample based on successive approximations of the target density π via a Gaussian kernel with precision matrix $\mathbf{P}$. Actually, searching the maximum a posteriori estimator under the proposal distribution κ in (4.31) boils down to solving

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\ \underbrace{-\log\pi(\boldsymbol{\theta})}_{f(\boldsymbol{\theta})} + \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)}\right\|_{\mathbf{P}}^2 \tag{4.47}$$

which coincides with step 4 in Algorithm 12 by taking $f = -\log\pi$. This puts a first emphasis on the tight connection between simulation and optimization that we already highlighted in previous sections.

4.4.2 *The G-PPA, ADMM and the approximate Richardson Gibbs sampler*

In the 1990's, the PPA has been used to improve existing optimization algorithms, namely Arrow-Hurwicz and Uzawa methods, to get rid of the assumption of strict convexity (Arrow, Hurwicz, and Uzawa, 1958; Bramble, Pasciak, and Vassilev, 1997; Ruszczynski, 1994; Kallio and Ruszczynski, 1994). More recently, an important motivation of the PPA has been related to the *preconditioning* idea used in the unifying model proposed

in (4.23). Indeed, the proximal mapping has been extensively used within the alternating direction method of multipliers (ADMM) (Glowinski and Marroco, 1975; Gabay and Mercier, 1976; Boyd et al., 2011) as a preconditioner in order to avoid high-dimensional inversions (Esser, Zhang, and Chan, 2010; Zhang, Burger, and Osher, 2011; Chambolle and Pock, 2011; Li, Sun, and Toh, 2016; Bredies and Sun, 2017). As introduced in Section 2.1.2 in Chapter 2, the ADMM stands for an optimization approach that solves the minimization problem in (4.45) when $f(\boldsymbol{\theta}) = f_1(\mathbf{H}\boldsymbol{\theta}) + f_2(\boldsymbol{\theta})$, $\mathbf{H} \in \mathbb{R}^{k \times d}$, via the following iterative scheme

$$\mathbf{z}^{(t)} = \underset{\mathbf{z} \in \mathbb{R}^k}{\arg\min}\ f_1(\mathbf{z}) + \frac{1}{2\rho} \left\| \mathbf{z} - \mathbf{H}\boldsymbol{\theta}^{(t-1)} - \mathbf{u}^{(t-1)} \right\|^2 \tag{4.48}$$

$$\boldsymbol{\theta}^{(t)} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\ f_2(\boldsymbol{\theta}) + \frac{1}{2\rho} \left\| \mathbf{H}\boldsymbol{\theta} - \mathbf{z}^{(t)} + \mathbf{u}^{(t-1)} \right\|^2 \tag{4.49}$$

$$\mathbf{u}^{(t)} = \mathbf{u}^{(t-1)} + \mathbf{H}\boldsymbol{\theta}^{(t)} - \mathbf{z}^{(t)}, \tag{4.50}$$

where $\mathbf{z} \in \mathbb{R}^k$ is a splitting variable, $\mathbf{u} \in \mathbb{R}^k$ is a scaled dual variable and ρ is a positive penalty parameter. One can notice that the $\boldsymbol{\theta}$-update (4.49) involves a matrix $\mathbf{H}$ operating directly on $\boldsymbol{\theta}$ preluding an expensive inversion of a high-dimensional matrix associated to $\mathbf{H}$. In addition, the presence of this matrix generally rules out the direct use of proximity operators to solve (4.49). To deal with such an issue, Algorithm 12 is considered to solve approximately the minimization problem in (4.49). The G-PPA applied to the minimization problem (4.49) reads

$$\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\ f_2(\boldsymbol{\theta}) + \frac{1}{2\rho} \left\| \mathbf{H}\boldsymbol{\theta} - \mathbf{z}^{(t)} + \mathbf{u}^{(t-1)} \right\|^2 + \frac{1}{2} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)} \right\|_{\mathbf{P}}^2. \tag{4.51}$$

With the particular choice $\mathbf{P} = \omega^{-1}\mathbf{I}_d - \rho^{-1}\mathbf{H}^T\mathbf{H}$, where $0 < \omega \leq \rho \left\|\mathbf{H}\right\|^{-2}$ ensures that $\mathbf{P}$ is positive semi-definite, the $\boldsymbol{\theta}$-update in (4.51) becomes

$$\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\ f_2(\boldsymbol{\theta}) + \frac{1}{2\omega} \left\| \boldsymbol{\theta} - \left(\boldsymbol{\theta}^{(t-1)} + \frac{\omega}{\rho}\mathbf{H}^T \left[\mathbf{z}^{(t)} - \mathbf{u}^{(t-1)} - \mathbf{H}\boldsymbol{\theta}^{(t-1)} \right] \right) \right\|^2. \tag{4.52}$$

This $\boldsymbol{\theta}$-update in (4.52) no more requires any computationally prohibitive matrix inversion but only matrix-vector products which are assumed to be undertaken efficiently without storing the whole corresponding matrix. Eventually, by defining $\mathbf{Q} \coloneqq \rho^{-1}\mathbf{H}^T\mathbf{H}$ and $\mathbf{P} = \omega^{-1}\mathbf{I}_d - \mathbf{Q}$, the application of the G-PPA to the ADMM can be seen as the deterministic equivalent of the approximate Richardson Gibbs sampler in Table 4.5. This highlights even more the tight links between (quadratic) optimization and Gaussian sampling. It also paves the way to novel sampling methods inspired by optimization approaches which are not necessarily dedicated to Gaussian sampling.

4.5 Conclusion

This chapter presented a general framework dedicated to high-dimensional Gaussian sampling. This framework has been shown to stand for a stochastic counterpart of the celebrated proximal point algorithm in optimization and as such could benefit from some of its numerous benefits (e.g., preconditioning of complicated quadratic potential functions). In addition, the proposed framework allowed to unify most of the existing MCMC methods which have been proposed to sample from high-dimensional Gaussian distributions. Beyond the unifying feature, we showed that this framework also permitted to relate and extend existing approaches.

Overall, this chapter shed light on two important aspects of this manuscript. Firstly, we showed that the potentially high-dimensional Gaussian distribution $\pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:b})$ in (3.26) which appears within SGS could be sampled efficiently. Combined with the fact that the conditional distributions $\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ were expected to be easy to sample from thanks to the AXDA framework, this chapter concluded demonstrating that the steps involved in SGS could be tackled efficiently. Finally, this chapter was in line with the results presented since the beginning of this manuscript by drawing another tight connection between the simulation and optimization fields.

5 Back to optimization: The tempered AXDA envelope

"Naturally there are side issues. To separate the main issue from the side issues is the first task of the orderly mind."

— Agatha Christie, *Dumb Witness*

Chapter contents

In Section 2.1.2 of Chapter 2, we already highlighted strong connections between common optimization schemes (e.g., quadratic penalty approaches) and SGS. In particular, we saw that the minimization steps involved in quadratic penalty methods corresponded to maximum a posteriori estimation problems under the full conditional distributions associated to the joint distribution $\pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:p})$ defined in (2.2). Similarly to recent works on proximal MCMC methods, this showed how a commonly-used optimization approach could be adapted to sample from a complicated probability distribution.

In this chapter, we propose to take the opposite path: starting from the marginal probability distribution $\pi_\rho(\boldsymbol{\theta})$ defined in (1.3), we will analyze the objective function that comes up when one wants to find the global maximum of $\pi_\rho(\boldsymbol{\theta})$. This analysis will shed new light on the approximation considered in the AXDA framework by focusing on potential functions and not on their respective probability distributions. Interestingly, we will see that the potential function of $\pi_\rho(\boldsymbol{\theta})$ defines a smooth approximation of the potential function f associated to the initial target distribution $\pi(\boldsymbol{\theta})$, and hence will be coined *envelope function* in relation to the works of Moreau (1965). This envelope function, which has been already used in the image processing and deep learning communities, will be shown to admit interesting regularity properties and strong relations with the Moreau envelope.

In Section 5.1, we present the so-called *tempered AXDA envelope*, standing for a tempered version of the potential function associated to $\pi_\rho(\boldsymbol{\theta})$, and the main motivations behind its analysis. Section 5.2 reviews existing works which either studied or used this envelope function. Finally, Section 5.3 derives the main properties of this envelope which are of interest if one wants to use the latter for optimization purposes. The results of this chapter are planned to be submitted before the end of the year.

5.1 The tempered AXDA envelope

After having motivated its analysis, this section defines the so-called *tempered AXDA envelope* which stands for a generalization of the potential function associated to the approximate marginal distribution $\pi_\rho(\boldsymbol{\theta})$ defined in (1.3).

5.1.1 *Motivations*

Up to now, we have tackled the problem of sampling from a complicated target probability distribution $\pi(\boldsymbol{\theta})$ by relying on an approximate statistical framework called AXDA, see Chapter 1. Nevertheless, although some works adopt a probabilistic (not necessarily Bayesian) approach and define a target probability distribution, they do not aim to sample from the latter but instead only seek to find its global mode. This is for instance the case for maximum likelihood estimation problems (Fadili and Bullmore, 2002; Idier, 2008; Filstroff, Lumbreras, and Févotte, 2018), or maximum a posteriori ones (Fadili and Starck, 2005; Pereyra et al., 2016) which boil down to the solving of a minimization problem. When working with $\pi_\rho(\boldsymbol{\theta})$, this minimization problem becomes

$$\text{Find } \boldsymbol{\theta}^\star \in \mathbb{R}^d \text{ such that } \boldsymbol{\theta}^\star \in \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\ f_\rho(\boldsymbol{\theta}) \neq \emptyset, \tag{5.1}$$

where the potential function f_ρ is defined by $f_\rho(\boldsymbol{\theta}) = -\log \pi_\rho(\boldsymbol{\theta})$ and has already been introduced in (1.31) in Chapter 1. In the previous chapters, few words have been said about the potential f_ρ and only a single approximation property has been shown, see Proposition 4 in Chapter 1. This property in particularly shows that f_ρ stands for an approximation of f, the potential function associated to the initial target distribution π in (1.1). In the sequel, we aim at providing further insights and quantitative properties about the approximate potential f_ρ to end up with a complete description of the approximation involved in the proposed AXDA framework, ranging from a simulation point of view to an optimization one.

5.1.2 *Definition*

As in Chapter 1, we will consider an extended real-valued function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ which verifies the following assumption:

$$f \text{ is proper, lower semi-continuous and coercive.} \tag{5.2}$$

We refer to its domain with the usual notation $\mathrm{dom} f = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid f(\boldsymbol{\theta}) < \infty\}$. In the following, our analysis will essentially focus on the particular case where $\kappa_\rho(\cdot;\boldsymbol{\theta}) = \mathcal{N}(\cdot;\boldsymbol{\theta}, \rho^2\mathbf{I}_d)$ which corresponds to performing a Gaussian smoothing of the initial distribution π. This choice is motivated by the simplicity induced by the Gaussian kernel, the fact that this kernel has been considered to derive quantitative bias and mixing time bounds in Chapters 2 and 3 and its link with the Moreau-Yosida regularization which

will be highlighted in Section 5.3. Under this assumption, the potential function f_ρ associated to the approximate marginal distribution π_ρ in (1.3) is defined, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$f_\rho(\boldsymbol{\theta}) = \frac{d}{2}\log(2\pi\rho^2) - \log \int_{\mathbb{R}^d} \exp\left(-\left[f(\mathbf{z}) + \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right]\right) \mathrm{d}\mathbf{z}. \quad (5.3)$$

This potential function can be interestingly seen as either a negative log-partition function associated to the conditional probability distribution $\pi_\rho(\mathbf{z}|\boldsymbol{\theta})$ or as a transform involving a so-called *Log-Int-Exp* (LIE) operator defined as

$$\mathrm{LIE}[g(\cdot)] = -\log \int_{\mathbb{R}^d} \exp[-g(\mathbf{z})]\mathrm{d}\mathbf{z}, \quad (5.4)$$

for functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} \exp[-g(\mathbf{z})]\mathrm{d}\mathbf{z} < \infty$. Adopting the latter point of view allows to relate the potential f_ρ to the discrete version of LIE, namely the celebrated *Log-Sum-Exp* (LSE) operator. This connection will allow to build on the already known properties of the LSE operator to derive interesting properties for f_ρ, see Sections 5.2 and 5.3. Taking inspiration from the LSE operator, we define hereafter a tempered version of the potential f_ρ in (5.3), coined *tempered AXDA envelope*.

Definition 1 (Tempered AXDA envelope)**.** *Let $\tau, \rho > 0$ stand for a temperature and a tolerance parameter, respectively. Assume that f satisfies (5.2). The tempered AXDA envelope of f is denoted $\mathcal{A}_\rho^\tau[f]$ and defined, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by*

$$\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) = \tau\,\mathrm{LIE}\left[\frac{f(\cdot) + \|\cdot - \boldsymbol{\theta}\|^2/(2\rho^2)}{\tau}\right] + d\alpha_\rho^\tau, \quad (5.5)$$

where $\alpha_\rho^\tau = \frac{\tau}{2}\log(2\pi\rho^2\tau)$. When the temperature is set to $\tau = 1$, the tempered AXDA envelope is denoted by $\mathcal{A}_\rho^1[f] := \mathcal{A}_\rho[f]$ and simply coined AXDA envelope.

Before showing the interesting properties of the tempered AXDA envelope, we will first further motivate its study by reviewing existing works that have considered and/or studied particular instances of this envelope.

5.2 Related works

This section further motivates the analysis of the tempered AXDA envelope by showing that the latter is strongly related to common approximate approaches (e.g., LSE and smoothing approaches in optimization) and has even been recently used to guide deep learning algorithms towards wide regions to improve their generalization properties.

5.2.1 *Log-Sum-Exp*

Contrary to the LIE operator (5.4) which can be associated to continuous distributions (e.g., propotional to $\exp[-g(\mathbf{z})]$ for $\mathrm{LIE}[g]$), the LSE operator is associated to categorical distributions over classes $i \in [d]$ with probability proportional to $\exp[-g(i)]$. For $g : [d] \to \mathbb{R}$, it is defined as

$$\mathrm{LSE}[g] = -\log \sum_{i=1}^{d} \exp[-g(i)]. \tag{5.6}$$

This operator has been extensively studied in the literature for two main reasons. First, the LSE operator is a standard way of performing a soft minimum, that is

$$\tau \,\mathrm{LSE}[g/\tau] \underset{\tau \to 0+}{\longrightarrow} \min_{i \in [d]} g(i), \tag{5.7}$$

where τ refers to a temperature parameter as in Definition 1. As such, it has been used as a smoothing approach in optimization, see for instance the recent review by Beck and Teboulle (2012) and references therein. The second interesting property of the LSE operator is commonly referred to as *Gumbel trick* in the machine learning community since it is related to the Gumbel distribution which appears in extreme value theory (Steutel and van Harn, 2003). This *trick* essentially allows to compute $\mathrm{LSE}[g]$ by perturbating the quantities $(g(i); i \in [d])$ with Gumbel random variables, taking their minimum component and finally their expectation, that is

$$\mathrm{LSE}[g] = \mathbb{E} \min_{i \in [d]} \{g(i) - G(i)\} + \gamma, \tag{5.8}$$

where $\gamma \approx 0.58$ is the Euler-Mascheroni constant and $G(i) \sim \mathrm{Gumbel}(0, 1)$ for $i \in [d]$. Thanks to this property, the Gumbel trick has been used to estimate log-partition functions via a so-called *perturb-and-MAP* approach (Papandreou and Yuille, 2011) and to sample from categorical distributions (Maddison, Tarlow, and Minka, 2014). Since the LIE operator stands for a continuous generalization of the LSE operator, we expect to retrieve similar properties as in the discrete case. The work by Maddison, Tarlow, and Minka (2014) partially answers this question by showing that the Gumbel trick can indeed be used to sample from continuous distributions on $\mathbb{R}^d$ via the use of a Gumbel process. In Section 5.3, we will use these results to show that the tempered AXDA envelope tends towards the Moreau envelope in the limiting case $\tau \to 0+$ and can be explicitly computed from the latter by using a Gumbel process.

Let $\mu > 0$ a location parameter and $\sigma > 0$ a scale parameter. The Gumbel distribution $\mathrm{Gumbel}(\mu, \sigma)$ is supported on $\mathbb{R}$ and admits the pdf $\theta \mapsto (1/\sigma)e^{-(u+e^{-u})}$ with $u = (\theta - \mu)/\sigma$. It is represented in Figure 5.1 for various tuples (μ, σ).

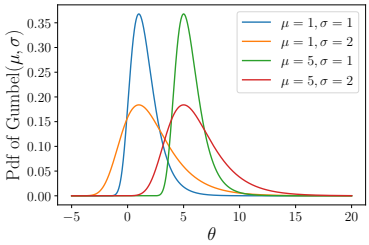


Figure 5.1: Probability density function associated to the Gumbel distribution $\mathrm{Gumbel}(\mu, \sigma)$.

5.2.2 *Smoothing envelopes*

In property *(iv)* of Proposition 1 (see Chapter 1), we showed that the approximate marginal distribution $\pi_\rho(\boldsymbol{\theta})$ is infinitely differentiable w.r.t. $\boldsymbol{\theta}$ for a Gaussian kernel κ_ρ. Since $\pi_\rho(\boldsymbol{\theta}) \propto \exp(-f_\rho(\boldsymbol{\theta}))$, this implies that the potential f_ρ in (5.3) is smooth and so is the tempered AXDA envelope $\mathcal{A}^\tau_\rho[f]$. Based on this smoothness property, we propose in this section to study the potential link between the tempered AXDA envelope

and the large family of smoothing approaches which are commonly employed in non-smooth optimization problems (Beck and Teboulle, 2012; Nesterov and Spokoiny, 2017). In few words, these smoothing methods transform a non-smooth optimization problem into an approximate smooth one by smoothing the non-smooth part of the former. In the convex scenario and by denoting $\epsilon > 0$ the prescribed precision on the objective function values, the main motivation is to build on faster gradient schemes which are shown to share an $\mathcal{O}(1/\epsilon)$ convergence rate compared to the slower $\mathcal{O}(1/\epsilon^2)$ rate associated to subgradient schemes. Beck and Teboulle (2012) proposed a unifying framework by defining the concept of *smoothable convex functions* which are functions that admit a smooth convex approximation. Interestingly, the LSE operator introduced in Section 5.2.1 fits within this framework. This framework encompasses also commonly-used smoothing approaches such as those using the Moreau envelope defined, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$\mathcal{M}_\rho[f](\boldsymbol{\theta}) = \inf_{\mathbf{z}\in\mathbb{R}^d} \left\{ f(\mathbf{z}) + \frac{1}{2\rho^2} \|\mathbf{z} - \boldsymbol{\theta}\|^2 \right\}, \tag{5.9}$$

or the Gaussian approximation by convolution (Nesterov and Spokoiny, 2017) defined by

$$\mathcal{C}_\rho[f](\boldsymbol{\theta}) = \int_{\mathbb{R}^d} f(\mathbf{z}) \frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{(2\pi\rho^2)^{d/2}} \mathrm{d}\mathbf{z}. \tag{5.10}$$

In Section 5.3, we will show that the tempered AXDA envelope shares similar regularity and approximation properties as the ones verified by the smoothing approaches detailed in (Beck and Teboulle, 2012) and hence stands for a surrogate smoothing method that can be resorted to in non-smooth optimization. In the three following sections, we do not only relate the proposed envelope to existing works but show that the tempered AXDA envelope $\mathcal{A}_\rho^\tau[f]$ has been already and explicitly used in the literature in various contexts.

5.2.3 *Local entropy*

We provide here a first particular context where the proposed AXDA envelope has been already and explicitly introduced, with some motivations far from the ones exposed in this thesis. In particular, in recent years, the AXDA envelope has shown to empirically improve the generalization properties of deep learning approaches.

In the seminal works by Baldassi et al. (2015), Baldassi et al. (2016), and Chaudhari et al. (2019), the AXDA envelope $\mathcal{A}_\rho[f]$, which equals the potential f_ρ in (5.3), has been coined *local entropy* and interestingly introduced within deep learning approaches. The aforementioned authors showed that the optimization (also called *energy* in statistical physics) landscape of neural networks is characterized by a lot of isolated and sharp minima, and by only a few dense regions gathering numerous and close local minima. In order to guide optimization algorithms towards

such wide valleys, they considered the minimization of $\mathcal{A}_\rho[f]$ instead of minimizing the initial non-convex loss function f.

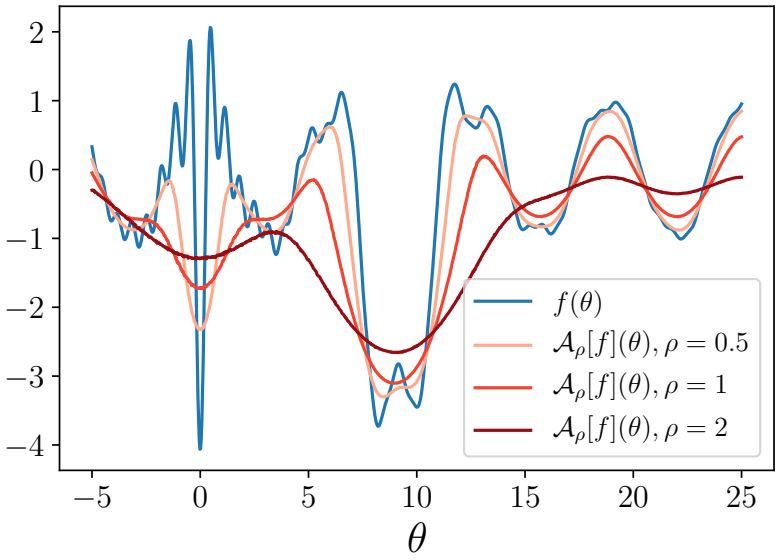


Figure 5.2: Behavior of the AXDA envelope $\mathcal{A}_\rho[f]$ w.r.t. ρ for complicated non-convex functions f.

Figure 5.2 illustrates such a complicated landscape in the univariate case and shows the landscapes associated to smooth approximations obtained via the AXDA envelope. To provide a simple and clear illustration, we constructed an initial non-convex function f with a very sharp global minimum at $\theta = 0$, lots of sharp local minima around this value and three wide valleys around $\theta = 9$, $\theta = 15$ and $\theta = 22$, respectively. One can denote that the AXDA envelope provides a way to guide optimization algorithms towards wide regions by concentrating on these regions and smoothing sharp and isolated local minima. As highlighted in Chapter 1 for $\pi_\rho(\boldsymbol{\theta})$, the larger ρ, the higher the smoothing. In Section 5.3, we will compare the smoothing induced by the AXDA envelope to the one associated to other smoothing approaches (e.g., the integral convolution of f or the Moreau envelope).

5.2.4 *Connections to Bayesian posterior mean estimators*

Both the AXDA envelope and its tempered version have been also recently considered in the image processing community to study Bayesian posterior mean estimators, also referred to as minimum mean square error (MMSE) estimators. Building upon the Tweedie's formula (Efron, 2011), Ong, Milanfar, and Getreuer (2019) for instance highlighted that the MMSE estimator associated to a denoising problem under Gaussian noise satisfies a shrinkage representation formula which involves the gradient of the AXDA envelope $\mathcal{A}_\rho[f]$.

In this section, we use these results to highlight complementary properties that are inherited by the proposed envelope. To this purpose, let consider the conditional probability density $\pi_\rho(\mathbf{z}|\boldsymbol{\theta})$ which appears when

considering an AXDA model (see Chapter 1) and writes

$$\pi_\rho(\mathbf{z}|\boldsymbol{\theta}) \propto \exp\left(-f(\mathbf{z}) - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right). \tag{5.11}$$

The mode of this distribution, denoted $\mathbb{M}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z)$, coincides with the proximity operator of f taken at $\boldsymbol{\theta}$, that is

$$\mathbb{M}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z) = \underset{\mathbf{z}\in\mathbb{R}^d}{\arg\min}\left\{f(\mathbf{z}) + \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right\} \tag{5.12}$$

$$\coloneqq \operatorname{prox}_{\rho^2 f}(\boldsymbol{\theta}). \tag{5.13}$$

If f is continuously differentiable and convex, then a characterization of the proximity operator of f can be obtained via first-order optimality conditions. More precisely, $\operatorname{prox}_{\rho^2 f}(\boldsymbol{\theta})$ can be shown to satisfy (via Fermat's rule):

$$\nabla_{\mathbf{z}}\left[f(\mathbf{z}) + \frac{\left\|\mathbf{z} - \operatorname{prox}_{\rho^2 f}(\boldsymbol{\theta})\right\|^2}{2\rho^2}\right] = \mathbf{0}_d. \tag{5.14}$$

After some calculus, this leads to the following property satisfied by the proximity operator of f:

$$\operatorname{prox}_{\rho^2 f}(\boldsymbol{\theta}) = (\mathbf{I} + \rho^2\nabla f)^{-1}(\boldsymbol{\theta}) \tag{5.15}$$

$$\underset{\rho\to 0}{\approx} \boldsymbol{\theta} - \rho^2\nabla f(\boldsymbol{\theta}). \tag{5.16}$$

This last line allows to interpret this mode as a shrinkage of $\boldsymbol{\theta}$ towards zero via the gradient of the potential f (Ong, Milanfar, and Getreuer, 2019). If we now look at the mean of Z under $\pi_\rho(\cdot|\boldsymbol{\theta})$, the latter can be interestingly interpreted as a shrinkage of $\boldsymbol{\theta}$ towards zero but via the gradient of $\mathcal{A}_\rho[f]$, that is (Ong, Milanfar, and Getreuer, 2019; Darbon and Langlois, 2020)

$$\mathbb{E}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z) = \boldsymbol{\theta} - \rho^2\nabla\mathcal{A}_\rho[f](\boldsymbol{\theta}). \tag{5.17}$$

In the scenario where f is convex, the connections between $\mathbb{E}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z)$ and the (tempered) AXDA envelope have been further studied in (Darbon and Langlois, 2020) where the authors in particular showed that the expectation $\mathbb{E}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z)$ satisfies the proximal mapping formula

$$\mathbb{E}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z) = \underset{\mathbf{z}\in\mathbb{R}^d}{\arg\min}\left\{\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2} + \left[\left(\frac{\|\cdot\|^2}{2} - \rho^2\mathcal{A}_\rho[f]\right)^*(\mathbf{z}) - \frac{\|\mathbf{z}\|^2}{2}\right]\right\}, \tag{5.18}$$

where g^* denotes the Fenchel-Legendre transform of the function g. Based on this representation, the expectation $\mathbb{E}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z)$ can be seen as the solution of a convex variational problem involving a Gaussian data fitting term and a regularization term. When tackling denoising problems, this finding can then be used to understand the properties of $\mathbb{E}_{\pi_\rho(\mathbf{z}|\boldsymbol{\theta})}(Z)$ which

Let $g : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ a proper, lower semi-continuous and convex function. The Fenchel-Legendre transform $g^* : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ of g is defined by

$$g^*(\boldsymbol{\theta}) = \sup_{\mathbf{u}\in\mathbb{R}^d}\left\{\mathbf{u}^T\boldsymbol{\theta} - g(\mathbf{u})\right\}.$$

is in this case an MMSE estimator, see the work by Darbon and Langlois (2020).

5.2.5 *Hamilton-Jacobi partial differential equation*

We finally briefly present a characterization of the tempered AXDA envelope that has been already highlighted in several works from distinct communities, see (Dolcetta, 2003; Chaudhari et al., 2018; Darbon and Langlois, 2020) and references therein.

Under mild assumptions on the potential f, these works pointed out that the tempered AXDA envelope $\mathcal{A}_\rho^\tau[f]$ stands for the unique viscosity solution to the Hamilton-Jacobi partial differential equation (PDE) defined by

$$\begin{cases} \frac{\partial u}{\partial \rho}(\boldsymbol{\theta}, \rho) + \frac{1}{2}\left\|\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta}, \rho)\right\|^2 = \frac{\tau}{2}\Delta_{\boldsymbol{\theta}} u(\boldsymbol{\theta}, \rho) \quad \text{for } (\boldsymbol{\theta}, \rho) \in \mathbb{R}^d \times (0, +\infty) \\ u(\boldsymbol{\theta}, 0) = f(\boldsymbol{\theta}) \quad \text{for } \boldsymbol{\theta} \in \mathbb{R}^d, \end{cases} \tag{5.19}$$

that is $u(\boldsymbol{\theta}, \rho) = \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta})$ satisfies (5.19). This characterization is particularly interesting since it is well-known that the Moreau envelope is the unique solution to the Hamilton-Jacobi PDE

$$\begin{cases} \frac{\partial u}{\partial \rho}(\boldsymbol{\theta}, \rho) + \frac{1}{2}\left\|\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta}, \rho)\right\|^2 = 0 \quad \text{for } (\boldsymbol{\theta}, \rho) \in \mathbb{R}^d \times (0, +\infty) \\ u(\boldsymbol{\theta}, 0) = f(\boldsymbol{\theta}) \quad \text{for } \boldsymbol{\theta} \in \mathbb{R}^d, \end{cases} \tag{5.20}$$

which again preludes the fact that the tempered AXDA envelope stands for a smooth proxy of the Moreau envelope and tends towards the latter as $\tau \to 0+$. This key property, among others, is enounced in the next section.

5.3 Properties

By reviewing some existing works related to the proposed envelope in Section 5.2, some nice and interesting properties of this envelope have been preluded such as its link to Moreau regularization or its smoothing properties. In this section, we propose to state and complement these properties in order to end up with a clear description of the tempered AXDA envelope and a fortiori of the approximation involved in the proposed AXDA framework.

5.3.1 *Standard properties*

We begin with standard results which are commonly expected when one wants to tackle the minimization problem in (5.1).

Proposition 10. *Let $\rho, \tau > 0$ and assume that f satisfies (5.2). Then, the following properties hold.*

(i) If $\mu(\mathrm{dom} f) > 0$ where $\mu(\cdot)$ is the d-dimensional Lebesgue measure, then $\mathcal{A}_\rho^\tau[f]$ is proper and $\mathrm{dom}\ \mathcal{A}_\rho^\tau[f] = \mathbb{R}^d$.

(ii) $\mathcal{A}_\rho^\tau[f]$ *is continuous and in particular lower semi-continuous.*

(iii) $\mathcal{A}_\rho^\tau[f]$ *is a coercive function, that is* $\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \to +\infty$ *as* $\|\boldsymbol{\theta}\| \to +\infty$.

(iv) If f *is (resp. strictly) convex, then* $\mathcal{A}_\rho^\tau[f]$ *is (resp. strictly) convex.*

Proof. See Appendix D.1. □

Combining properties (i), (ii) and (iii) of this proposition, it follows from the Bolzano-Weierstrass theorem that $\mathcal{A}_\rho^\tau[f]$ admits a unique minimum $\boldsymbol{\theta}^\star$ on $\mathbb{R}^d$ (Beck, 2017, Chapter 2). When the function f is in addition strictly convex and satisfies a symmetry property, Proposition 11 shows that the global minimum $\boldsymbol{\theta}^\star$ of $\mathcal{A}_\rho^\tau[f]$ coincides with the unique minimum of f.

Proposition 11. *Assume that* $\mathcal{A}_\rho^\tau[f]$ *admits a global minimum* $\boldsymbol{\theta}^\star$*, and that* f *is strictly convex and admits a unique minimum* $\boldsymbol{\theta}_0$ *such that for all* $\boldsymbol{\theta} \in \mathbb{R}^d$, $f(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) = f(\boldsymbol{\theta}_0 + \boldsymbol{\theta})$*. Then,* $\boldsymbol{\theta}^\star$ *is the unique minimum of* $\mathcal{A}_\rho^\tau[f]$ *and verifies* $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_0$.

Proof. See Appendix D.2. □

5.3.2 *Approximation and smoothing properties*

In the following, we will show that $\mathcal{A}_\rho^\tau[f]$ can, in some sense, be related to ρ-smooth approximations of f over $\mathrm{dom} f$ as defined in (Beck and Teboulle, 2012, Definition 2.1). This definition is recalled hereafter for completeness.

Definition 2. *Let* f *satisfy* (5.2) *and* $\Theta \subseteq \mathrm{dom} f$ *be a closed set. The function* f *is called* (α, β)*-smoothable over* Θ *if there exist* β_1, β_2 *satisfying* $\beta_1 + \beta_2 = \beta$ *such that for any* $\rho > 0$ *there exists a continuously differentiable function* $\tilde{f}_\rho : \mathbb{R}^d \to (-\infty, +\infty)$ *such that the following hold:*

(i) $-\beta_1\rho \leq \tilde{f}_\rho(\boldsymbol{\theta}) - f(\boldsymbol{\theta}) \leq \beta_2\rho$*, for every* $\boldsymbol{\theta} \in \Theta$.

(ii) $\tilde{f}_\rho$ *has a* α/ρ*-Lipschitz gradient over* Θ.

The function $\tilde{f}_\rho$ *is called a* ρ*-smooth approximation of* f *over* Θ *with parameters* (α,β).

Property (i) simply states that $\tilde{f}_\rho$ is an arbitrary tight approximation of the original function f and that its bias is controlled by a tolerance parameter ρ. On the other hand, property (ii) ensures that $\tilde{f}_\rho$ is smooth, i.e., it is continuously differentiable and admits a Lipschitz-continuous gradient function. Its associated Lipschitz constant is inversely proportional to ρ which encodes the fact that $\tilde{f}_\rho$ is smoother as ρ increases. Combining these two properties, one can denote that the tolerance parameter ρ stands for a trade-off between accuracy and smoothness.

We already showed that $\mathcal{A}_\rho^\tau[f]$ is continuously differentiable (see property (iv) of Proposition 1) and takes its values in $\mathbb{R}$ (see property (i) of Proposition 10). When f is Lipschitz-continuous and $\mathrm{dom} f = \mathbb{R}^d$, we also showed in Proposition 4 that there exist two constants $c_1(\rho)$ and $c_2(\rho)$ satisfying $c_{i,i\in\{1,2\}} = \mathcal{O}(\rho)$ for sufficiently small ρ and such that for all $\boldsymbol{\theta} \in \mathrm{dom} f$,

$$c_1(\rho) \leq \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) - f(\boldsymbol{\theta}) \leq c_2(\rho).$$

This last property is obviously weaker than (i) in Definition 2 since the linear approximation in ρ only holds for sufficiently small values of ρ. Nevertheless, it still shows that $\mathcal{A}_\rho^\tau[f]$ tends towards f when ρ tends towards 0^+. We now prove that $\mathcal{A}_\rho^\tau[f]$ is gradient-Lipschitz with a Lipschitz constant of the form α/ρ^2 with $\alpha > 0$. This property is stated in the following proposition where we show that $\mathcal{A}_\rho^\tau[f]$ is smooth without requiring the convexity of f compared to inf-conv smooth approximations (Moreau, 1965; Beck and Teboulle, 2012).

Proposition 12. *Let $\tau, \rho > 0$ and assume that f satisfies (5.2). Then, the following properties hold.*

(i) If f is L-Lipschitz continuous, then $\mathcal{A}_\rho^\tau[f]$ is L_ρ-Lipschitz continuous with $L_\rho \leq L$.

(ii) $\mathcal{A}_\rho^\tau[f]$ is continuously differentiable and admits a M_ρ-Lipschitz continuous gradient with $M_\rho \leq 1/\rho^2$. For all $\boldsymbol{\theta} \in \mathbb{R}^d$, its gradient writes

$$\nabla \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) = \frac{\boldsymbol{\theta} - \mathbb{E}_{\pi_\rho^\tau(\mathbf{z}|\boldsymbol{\theta})}(Z)}{\rho^2}, \tag{5.21}$$

where

$$\pi_\rho^\tau(\mathbf{z}|\boldsymbol{\theta}) \propto \exp\left(-\frac{f(\mathbf{z})}{\tau} - \frac{\|\boldsymbol{\theta} - \mathbf{z}\|^2}{2\rho^2\tau}\right). \tag{5.22}$$

Proof. See Appendix D.3. □

Beyond the fact that $\mathcal{A}_\rho^\tau[f]$ is indeed smooth, one can denote that its gradient involves the mean under a tempered version of the conditional distribution $\pi_\rho(\mathbf{z}|\boldsymbol{\theta})$, which has been studied in the recent work by Darbon and Langlois (2020). This property will in particular suggest the definition of a smooth proximity operator, see Definition 3 in Section 5.3.4.

5.3.3 *A compromise between integral and infimal convolutions*

As pointed out in Section 5.2.2, multiple smoothing methods have been already proposed to ease and/or accelerate the convergence of optimization algorithms. In this section, we compare the proposed envelope to two classical and highly-used smooth approximations, namely the Moreau envelope and the Gaussian approximation by convolution, introduced in Section 5.2.2. In the convex case, the following proposition shows that the AXDA tempered envelope lies in between the two aforementioned approximations.

Proposition 13. *Let $\tau, \rho > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ and assume that f satisfies (5.2). The following inequalities hold.*

(i) $\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \leq \mathcal{C}_{\rho\sqrt{\tau}}[f](\boldsymbol{\theta})$.

(ii) If f is convex, then $\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \geq \mathcal{M}_\rho[f](\boldsymbol{\theta})$.

Proof. See Appendix D.4. □

Figure 5.3 illustrates Proposition 13 by considering one convex function (the absolute loss) and a non-convex one (an indicator function over $[-1, 1]$). For simplicity, we set $\rho = \tau = 1$. For these two functions, the three envelope functions that are considered admit a closed-form expression, see for instance (1.53) in Chapter 1 for $\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta})$ with $f = |\cdot|$. In this figure, one can denote that the tempered AXDA envelope indeed lies in between the Moreau and integral convolution envelopes for the convex absolute loss. On the other hand, for the characteristic function over $[-1, 1]$, only the inequality (i) relating $\mathcal{A}_\rho^\tau[f]$ and $\mathcal{C}_{\rho\sqrt{\tau}}[f]$ holds. For this non-convex function, one can also note that the tempered AXDA envelope is smoother than the Moreau one, see property (ii) of Proposition 12.

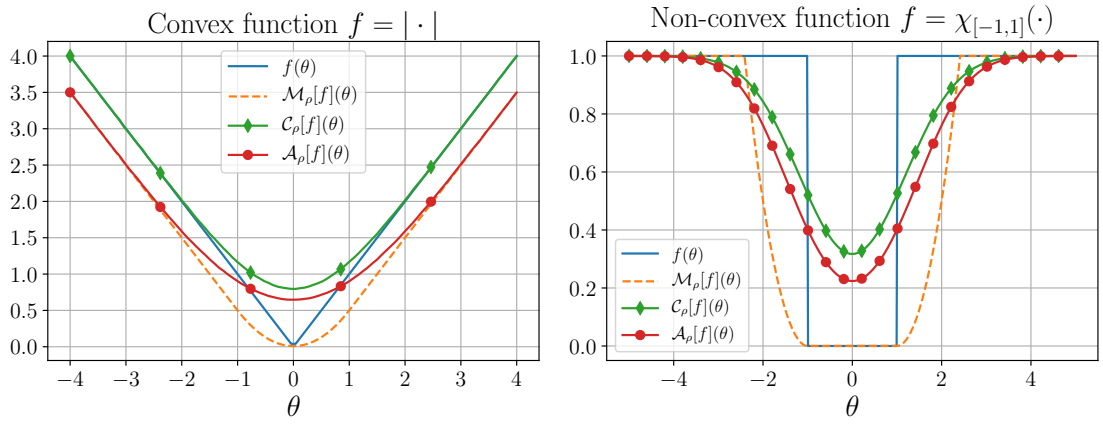


Figure 5.3: Illustrations of Proposition 13 with $\rho = \tau = 1$; (left) $f(\boldsymbol{\theta}) = |\boldsymbol{\theta}|$; (right) $f(\boldsymbol{\theta}) = \chi_{[-1,1]}(\boldsymbol{\theta})$, i.e., $f(\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \in [-1, 1]$ and $f(\boldsymbol{\theta}) = 1$ (instead of ∞) otherwise.

5.3.4 *Explicit relations with the Moreau envelope and the proximity operator*

Expected relations with the Moreau envelope – In Section 5.2, we already provided some insights which preluded the strong relationship between the proposed tempered AXDA envelope and the Moreau envelope. In the following proposition, we confirm these findings by showing that the tempered AXDA envelope not only stands for a smooth approximation of f (via ρ) but is also a smooth proxy of the associated Moreau envelope (via the temperature τ).

Proposition 14. *Let $\rho, \tau > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. Assume that f is proper, convex and lower semi-continuous, and such that* $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$. *Then* $\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \to \mathcal{M}_\rho[f](\boldsymbol{\theta})$ *as* $\tau \to 0^+$.

Proof. See (Darbon and Langlois, 2020, Theorem 3.1). □

Figure 5.4 illustrates this convergence property with the absolute loss function considered in Figure 5.3.

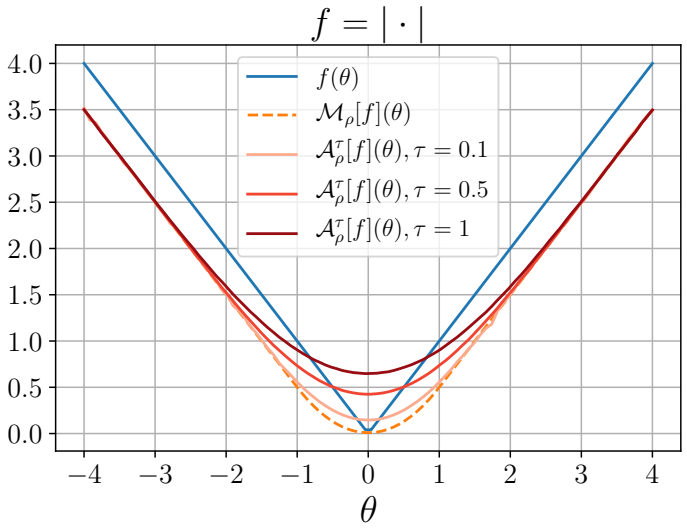


Figure 5.4: Illustration of Proposition 14 with $\rho = 1$ and $f(\boldsymbol{\theta}) = |\boldsymbol{\theta}|$.

Similarly to the properties of the LSE operator detailed in Section 5.2.1, the next proposition shows that the AXDA envelope of a function f can be explicitly computed from a perturbed version of the Moreau envelope of f via a Gumbel process.

Proposition 15. *Let $\rho > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. Then, we have:*

$$\mathcal{A}_\rho[f](\boldsymbol{\theta}) = \mathbb{E}(\mathcal{M}_\rho[f - G](\boldsymbol{\theta})) + \gamma, \tag{5.23}$$

where $\forall \mathbf{z} \in \mathbb{R}^d$, $G(\mathbf{z}) \overset{iid}{\sim} \mathrm{Gumbel}(0, 1)$ and $\gamma \approx 0.58$ is the Euler-Mascheroni constant.

Proof. The proof follows from the continuous version of the Gumbel trick, see the work by Maddison, Tarlow, and Minka (2014). □

As for the LSE operator, this property allows to interpret the proposed envelope as an expectation of a perturbed minimum and to relate it to perturb-and-MAP approaches. As an example, in the work by Maddison, Tarlow, and Minka (2014), this continuous generalization of the celebrated Gumbel trick has been used to sample from a complicated target distribution by converting this sampling task into an optimization problem.

The soft-proximity operator – The Moreau envelope being related to the proximity operator, we seize the opportunity of studying the tempered AXDA envelope to introduce a so-called *τ-soft proximity operator* which stands for an arbitrary tight and smooth approximation of the proximity operator. In (5.21), we saw that the gradient of the tempered AXDA envelope satisfies a relation involving the mean under the conditional density $\pi_\rho(\mathbf{z}|\boldsymbol{\theta})$. This relation can be compared to the well-known property

satisfied by the gradient of the Moreau envelope, that is

$$\nabla \mathcal{M}_\rho[f](\boldsymbol{\theta}) = \frac{\boldsymbol{\theta} - \operatorname{prox}_{\rho^2 f}(\boldsymbol{\theta})}{\rho^2}. \tag{5.24}$$

By identification, we propose the following approximation for the proximity operator.

Definition 3. *Let $\rho, \tau > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. The so-called τ-soft proximity operator associated to f is defined by*

$$\operatorname{sprox}_f^{\rho^2,\tau}(\boldsymbol{\theta}) = \boldsymbol{\theta} - \rho^2 \nabla \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \tag{5.25}$$

$$= \mathbb{E}_{\pi_\rho^\tau(\mathbf{z}|\boldsymbol{\theta})}(Z), \tag{5.26}$$

where $\pi_\rho^\tau(\mathbf{z}|\boldsymbol{\theta})$ has been defined in (5.22).

The following proposition shows that this operator is continuous and indeed tends towards the celebrated proximity operator as the temperature parameter τ vanishes.

Proposition 16. *Assume that f is proper, convex and lower semi-continuous, and such that $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$. The τ-soft proximity operator associated to the function f satisfies the following properties.*

(i) The τ-soft proximity operator is continuous w.r.t. τ on $(0, \infty)$.

(ii) Let $\boldsymbol{\theta} \in \mathbb{R}^d$. If f is convex, then $\operatorname{sprox}_f^{\lambda,\tau}(\boldsymbol{\theta}) \to \operatorname{prox}_{\lambda f}(\boldsymbol{\theta})$ as $\tau \to 0^+$.

Proof. Property (i) follows from the continuity theorem under the integral sign. The proof of property (ii) can be found in (Darbon and Langlois, 2020, Theorem 3.1). □

When the proximity operator of a given convex function f is not easily available, this soft-proximity operator might be a way to compute an approximate proximity operator by sampling from the tempered conditional distribution $\pi_\rho^\tau(\mathbf{z}|\boldsymbol{\theta})$. As an example, Chaudhari et al. (2019) built on Langevin dynamics to estimate $\operatorname{sprox}_f^{\lambda,\tau}(\boldsymbol{\theta})$. Obviously, some approximation guarantees are required in order to ensure that the proposed soft proximity operator is sufficiently close to the actual proximity operator. When f is m-strongly convex, Darbon and Langlois (2020) recently showed that the mean square error between $\operatorname{sprox}_f^{\lambda,\tau}(\boldsymbol{\theta})$ and $\operatorname{prox}_{\lambda f}(\boldsymbol{\theta})$ is of the order $\mathcal{O}(\tau)$.

5.4 Conclusion

This chapter concluded the analysis of the proposed AXDA framework by focusing on the potential function f_ρ associated to the marginal distribution $\pi_\rho(\boldsymbol{\theta})$. Interestingly, we saw that this interesting potential function admitted better properties than the initial potential function f such as

smoothness. As such, it has been both used and studied in the recent literature. Capitalizing on its discrete counterpart involving the LSE operator, we showed that a tempered version of f_ρ, coined tempered AXDA envelope, could be seen as a smooth approximation of the celebrated Moreau envelope. This connection permitted, in particular, to compare the approximation involved in the AXDA framework to the one used in proximal MCMC methods. The study of this envelope function ended by the introduction of an approximate proximity operator, recently studied in (Darbon and Langlois, 2020).

The results of this chapter showed that the proposed AXDA framework yields a family of envelope functions that can be used in optimization. Interestingly, all the work presented in this manuscrit could have been derived in the opposite way. Indeed, we could have started from Chapter 5 with the study of the proposed tempered envelope function for optimization purposes before using it to define a family of approximate densities π_ρ, see Chapter 1. This shows the generality and the multiple dimensions of this work which obviously open new research routes as shown in the general Conclusion section below.

Conclusion

This thesis has developed a generic approximate statistical framework, coined AXDA, for infering unknown parameters in complex and high-dimensional models. The proposed approach defined a systematic approximate data augmentation scheme. By targeting this augmented model with a Gibbs sampler, the proposed AXDA framework permitted to fulfill important requirements: (i) simple inference through a divide-to-conquer scheme, (ii) theoretical guarantees, (iii) scalable MCMC sampling in both high-dimensional and distributed settings and (iv) strong relationships with optimization opening new research routes. Apart from these properties, AXDA appeared to be a unifying and rich class of models with numerous interpretations, both from a simulation and an optimization point of view.

A general, rich and unifying framework for statistical inference

Chapter 1 presented the proposed approximate statistical framework and its main ingredient, namely a Dirac-delta converging sequence ensuring the recovery of the initial target density in a limiting case. We showed how to build this sequence in a systematic manner leading to a widely applicable framework compared to case-specific data augmentation schemes. By reviewing existing related approaches, we identified general properties and algorithms that could be inherited by AXDA (e.g, robustness or sophisticated schemes from the ABC literature). We finally derived numerous non-asymptotic theoretical guarantees permiting to assess the bias of AXDA in different scenarios. These results have been submitted to an international journal (Vono, Dobigeon, and Chainais, 2020a) and presented in a national conference (Vono, Dobigeon, and Chainais, 2019d).

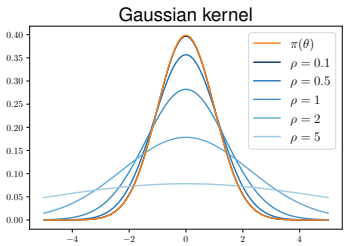


Figure 5.5: Illustration of the proposed approximation built with a Gaussian kernel $\mathcal{N}(0, \rho^2)$ when the target distribution is Gaussian.

Chapter 2 presented how to perform a simple, scalable and efficient inference thanks to a Gibbs sampler, coined SGS, targeting an AXDA model. We showed that SGS could be interestingly seen as a stochastic counterpart of quadratic penalty methods and benefited from the same important advantage as ADMM, namely the division of a difficult problem into simpler ones. These benefits have been illustrated on challenging Bayesian inference problems where SGS successfully managed to generate samples from complicated target posterior distributions in a reasonable amount of time. It was also able to tackle efficiently difficult problems that cannot be directly addressed with state-of-the-art methods, such as Poissonian image restoration. These results have been published in an international journal (Vono, Dobigeon, and Chainais, 2019a) and pre-

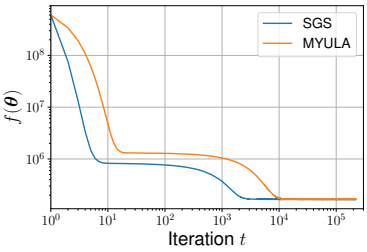


Figure 5.6: Convergence to the typical set of the posterior distribution π for SGS and MYULA.

sented in both international and national conferences (Vono, Dobigeon, and Chainais, 2019c; Vono, Dobigeon, and Chainais, 2018; Vono, Dobigeon, and Chainais, 2019b, 2019e).

Chapter 3 derived explicit, non-asymptotic and dimension-free convergence rates for SGS under classical assumptions that could be verified in practice. By combining these results with the bias bounds shown in Chapter 1, we showed complexity results for SGS which admit explicit dependencies with respect to the dimension of the problem, the prescribed precision and regularity constants associated to the target posterior distribution. In the single splitting scenario, we showed that these results improved upon those that have been proven so far for classical MCMC schemes (e.g., ULA and HMC). The results of this chapter provided again evidences that the proposed AXDA framework was able to yield scalable sampling approaches. These results have been submitted to an international journal (Vono, Paulin, and Doucet, 2019).

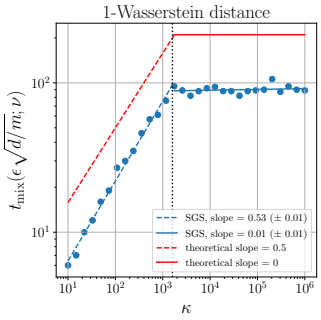


Figure 5.7: Multivariate Gaussian. $\epsilon\sqrt{d/m}$-mixing times for the 1-Wasserstein distance.

Chapter 4 complemented these evidences by showing that the Gaussian sampling step within SGS could be addressed efficiently by using state-of-the-art techniques. In this chapter, we did not only review such sampling techniques but proposed also to shed new light on these methods by embedding them into a unifying framework based on a stochastic version of the celebrated proximal point algorithm. Similarly to the connections between SGS and quadratic penalty methods, this framework permited to demonstrate new sampling routes by exploiting ideas from the optimization community. These results have been submitted to an international journal (Vono, Dobigeon, and Chainais, 2020b).

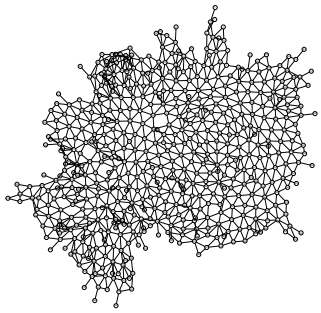


Figure 5.8: Graph defined on the 544 regions of Germany.

Chapter 5 finally presented a complementary interpretation of the approximation involved in the proposed AXDA framework. By focusing on the potential function f_ρ associated to the approximate marginal density built via AXDA, we showed that this function admitted useful properties and as such has been used in many applications. On top of these properties, we showed that the AXDA approximation could be explicitly related to the approximation associated to proximal MCMC schemes. Indeed, a tempered version of the potential f_ρ has been shown to converge towards the celebrated Moreau envelope in optimization. This chapter again contributed to highlight the tight link between optimization and sampling that has been at the heart of this manuscript.

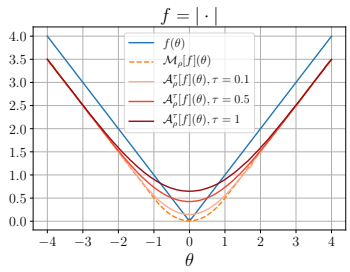


Figure 5.9: Convergence of the proposed tempered envelope $\mathcal{A}^\tau_\rho[f]$ towards the Moreau envelope $\mathcal{M}_\rho[f]$, with $\rho = 1$ and $f(\boldsymbol{\theta}) = |\boldsymbol{\theta}|$.

For sake of reproducible research, the code associated to the numerical results presented in our research works is available online at

https://github.com/mvono

Perspectives and future works

The generality of the proposed AXDA framework and associated MCMC sampling algorithm, along with their strong relations with optimization, allow to consider various prospective works. They can be divided into

methodological generalizations, theoretical contributions and new applications.

Methodological generalizations

Adaptive SGS sampling – So far, the presentation and the application of the AXDA framework has considered a fixed tolerance parameter ρ. In Chapters 2 and 3, we saw that the performance of SGS via its mixing properties is highly sensitive to this parameter: a large value yielded fast convergence but high bias while a small value deteriorated the mixing properties of the Markov chain. In Chapter 3, we already gave explicit guidelines to choose ρ based on the prescribed precision ϵ which upper-bounded a given statistical distance between the true target π and the approximate one π_ρ. On the other hand, Rendell et al. (2018) proposed an optimal selection of this parameter in a simple univariate Gaussian scenario, and its adaptive selection within a sequential Monte Carlo algorithm targeting π_ρ. Nevertheless, it is still unknown at the moment whether an adaptive SGS associated to a sequence $\{\rho_k\}_{k\in\mathbb{N}}$ will perform better than its standard version with a fixed tolerance parameter ρ. Hence, an interesting methodological extension of the proposed work is to derive an adaptive SGS sampling strategy which will permit to bypass the empirical tuning of ρ.

(A)synchronous SGS sampling – In Chapter 1, we highlighted the fact that the proposed AXDA framework permitted to scale MCMC sampling algorithms, and in particularly SGS, to distributed environments. However, even if a distributed SGS is highly attractive for solving big data problems, it suffers from the issue that all the auxiliary variables have to be synchronized to sample the master variable $\boldsymbol{\theta}$. This synchronization constraint becomes problematic if the different workers have different delays (e.g., due to different processing units) and if the sampling difficulty associated to each conditional probability distribution $\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ differs. Indeed, in this case, one has to wait for the slowest worker to update the master variable which might severely slow down the sampling procedure. To cope with these issues, an alternative might be to derive an asynchronous version of SGS.

Primal-dual sampling – In Chapter 2, we saw that the proposed SGS was the stochastic sampling counterpart of quadratic penalty methods which approximately solve a complicated optimization problem. A way to correct this approximation while still benefiting from its advantages (e.g., simple minimization steps) is to use the celebrated ADMM approach which introduces dual variables and invokes the duality principle in optimization. This triggers the very natural question: Is there an equivalent of this duality principle for sampling? Up to our knowledge, this *statistical* duality principle concept and its link with the common duality in optimization are not clear at the moment. We strongly believe that this research route is worth studying and can extend the results presented in this manuscript. Indeed, ADMM, which can be related to SGS in some sense, is known to be a special instance of so-called first-order primal-dual algorithms

(Chambolle and Pock, 2011). Hence, answering the above question might allow to derive stochastic counterparts of these optimization algorithms and contribute to further understand the relations between simulation and optimization.

Theoretical contributions

Non-asymptotic analysis of SGS with partial splitting – When the potential function associated to the target distribution can be written as $\sum_{i=1}^{b} f_i$, the non-asymptotic analysis of SGS presented in Chapter 3 focused on the specific case where all the individual contributions f_i have been *split*. However in some applications, one does not split all these contributions but only a fraction of it leading to a partial splitting strategy, see for instance the inpainting application in Chapter 2. For those cases, the theory developed in Chapter 3 cannot be directly used. Hence, another possible extension of this work is to derive explicit convergence rates for SGS under a partial splitting strategy.

Study of the soft-proximity operator – In Chapter 5, we introduced the so-called *soft-proximity operator* which stands for an approximation of the proximity operator in optimization. As pointed out in this chapter, some properties of the soft-proximity operator have already been derived in the works by Ong, Milanfar, and Getreuer (2019) and Darbon and Langlois (2020) in the Bayesian scenario where the likelihood is a Gaussian with diagonal covariance matrix and the prior is log-concave. Nevertheless, these works did not analyze this operator from an optimization point of view. They rather chose to investigate it from a Bayesian perspective (where it actually stands for an MMSE estimator) and compared it to the MAP estimator. Given the large impact the proximity operator had in the optimization literature, we strongly believe that this soft-proximity operator still admits unexplored properties that might be of interest in optimization. For instance, it is not clear at the moment if the soft-proximity operator enjoys a sort of Moreau decomposition formula and how it relates to the notion of projection. Hence, an interesting extension of this work could be the theoretical study of this soft-proximity operator from an optimization perspective.

Applications to other challenging problems

Probability densities that are not log-concave – We conclude on the potential prospective works associated to this manuscript by pointing out some applied problems where the proposed AXDA methodology has not been applied so far, namely Bayesian inference problems where the posterior is not log-concave and potentially multimodal. For such problems, the efficiency of SGS is unknown and might deteriorate due to the multimodality of the target and the sequential nature of the Gibbs sampling procedure. A typical example is the blind source separation and its constrained formulations (nonnegative matrix factorization, linear unmixing) which are ubiquituous in various applicative domains such as astrophysics

and hyperspectral imaging (Bobin et al., 2008). This problem involves the joint estimation of the mixing matrix and the sources, and yields a high-dimensional posterior distribution which is not log-concave. Hence, a possible research route could be to analyze if the proposed MCMC sampling algorithm can efficiently tackle such problem and indeed scale with the dimension of the state space.

Appendices – Chapter 1

Chapter contents

A.1 Proof of Proposition 1

Property *(i)* follows from the fact that π_ρ stands for a convolution integral between π and κ_ρ, i.e., $\pi_\rho = \pi * \kappa_\rho$. Therefore, the expectation and variance under π_ρ are the sum of the expectations and variances of two independent random variables under π and κ_ρ respectively.

Property *(ii)* follows directly from Folland (1999, Proposition 8.6).

Property *(iii)* follows from the fact that log-concavity is preserved by marginalization (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).

A.2 Proof of Proposition 2

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

$$
\begin{aligned}
W_p^p(\pi, \pi_\rho) &= \min_{\mu \in \Gamma(\pi_\rho, \pi)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|^p \, \mathrm{d}\mu(\boldsymbol{\theta}, \mathbf{z}) && \text{(A.1)} \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|^p \, \pi_\rho(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{z} && \text{(A.2)} \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|^p \, \kappa_\rho(\mathbf{z}; \boldsymbol{\theta}) \pi(\mathbf{z}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{z} && \text{(A.3)} \\
&= \rho^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \mathbf{z}\|^p \, K(\rho^{-1}(\boldsymbol{\theta} - \mathbf{z})) \pi(\mathbf{z}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{z} && \text{(A.4)} \\
&= \rho^p \int_{\mathbb{R}^d} \|\mathbf{u}\|^p \, K(\mathbf{u}) \mathrm{d}\mathbf{u} \int_{\mathbb{R}^d} \pi(\mathbf{z}) \mathrm{d}\mathbf{z} && \text{(A.5)} \\
&= \rho^p \int_{\mathbb{R}^d} \|\mathbf{u}\|^p \, K(\mathbf{u}) \mathrm{d}\mathbf{u}. && \text{(A.6)}
\end{aligned}
$$

A.3 Proof of Proposition 3

Let $\boldsymbol{\theta} \in \mathbb{R}^d$. Since π 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

$$\pi_\rho(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} \pi(\mathbf{z})\kappa_\rho(\mathbf{z};\boldsymbol{\theta})\mathrm{d}\mathbf{z} \tag{A.7}$$

$$= \frac{\displaystyle\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}) \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}} \tag{A.8}$$

$$= \frac{\displaystyle\int_{\mathbb{R}^d} \left[\pi(\boldsymbol{\theta}) - \sqrt{\rho}\nabla\pi(\boldsymbol{\theta})^T\mathbf{u} + \frac{\rho}{2}\mathbf{u}^T\mathbf{H}_\pi(\tilde{\boldsymbol{\theta}})\mathbf{u}\right] \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}}, \tag{A.9}$$

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

It follows that

$$\pi_\rho(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \tag{A.10}$$

$$- \sqrt{\rho}\nabla\pi(\boldsymbol{\theta})^T \int_{\mathbb{R}^d} \mathbf{u} \frac{\displaystyle\exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}} \tag{A.11}$$

$$+ \frac{\rho}{2} \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{u}^T\mathbf{H}_\pi(\tilde{\boldsymbol{\theta}})\mathbf{u} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta})}{\rho}\right) \mathrm{d}\mathbf{u}} \tag{A.12}$$

We now show that (A.11) $= O(\sqrt{\rho})$ and (A.12) $= \mathcal{O}(\rho)$. To this purpose, we use the analyticity and two times differentiability of d_ψ w.r.t. to its first argument and the continuity of $\mathbf{H}_{d_\psi}$. By definition of the Bregman divergence, $d_\psi(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$ and $\nabla_{\mathbf{z}} d_\psi(\mathbf{z}, \boldsymbol{\theta})\Big|_{\mathbf{z}=\boldsymbol{\theta}} = \mathbf{0}_d$ so that, for all $\mathbf{u} \in \mathbb{R}^d$,

$$d_\psi(\boldsymbol{\theta} - \sqrt{\rho}\mathbf{u}, \boldsymbol{\theta}) = \frac{\rho}{2}\mathbf{u}^T\mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}, \tag{A.13}$$

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

We first prove (A.12) $= \mathcal{O}(\rho)$. Using (A.13), we can re-write (A.12) as

$$(12) = \frac{\rho}{2} \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{u}^T\mathbf{H}_\pi(\tilde{\boldsymbol{\theta}})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) \mathrm{d}\mathbf{u}}. \tag{A.14}$$

Since $\lim_{\rho\to 0} \boldsymbol{\theta}' = \boldsymbol{\theta}$ and $\lim_{\rho\to 0} \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$, we will use the dominated

convergence theorem using that

$$\lim_{\rho\to 0} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) = \mathbf{u}^T \mathbf{H}_\pi(\boldsymbol{\theta})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right). \tag{A.15}$$

By using that $\|\mathbf{H}_\pi\| \leq C < \infty$ and $\left\|\mathbf{H}_{d_\psi}\right\| \geq c > 0$, we have:

$$\left|\mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right)\right| \leq C \|\mathbf{u}\|^2 \exp\left(-\frac{c}{2}\mathbf{u}^T\mathbf{u}\right), \tag{A.16}$$

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

$$\int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\tilde{\boldsymbol{\theta}})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) \mathrm{d}\mathbf{u} = \int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\boldsymbol{\theta})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u} + o(1). \tag{A.17}$$

Similarly,

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) \mathrm{d}\mathbf{u} = \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u} + o(1). \tag{A.18}$$

Hence,

$$(12) = \frac{\rho}{2} \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{u}^T \mathbf{H}_\pi(\boldsymbol{\theta})\mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u}} + o(\rho) \tag{A.19}$$

$$= \frac{\rho}{2} \mathrm{Trace}\left(\mathbf{H}_\pi(\boldsymbol{\theta}) \mathbf{H}_{d_\psi}(\boldsymbol{\theta})^{-1}\right) + o(\rho). \tag{A.20}$$

We now prove (A.11) $= \mathcal{O}(\sqrt{\rho})$. Using (A.13), it follows that

$$(11) = -\sqrt{\rho} \nabla\pi(\boldsymbol{\theta})^T \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta}')\mathbf{u}\right) \mathrm{d}\mathbf{u}}. \tag{A.21}$$

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

$$(11) = -\sqrt{\rho} \nabla\pi(\boldsymbol{\theta})^T \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u}} + o(\sqrt{\rho}) \tag{A.22}$$

$$= -\sqrt{\rho} \nabla\pi(\boldsymbol{\theta})^T \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{u} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{H}_{d_\psi}(\boldsymbol{\theta})\mathbf{u}\right) \mathrm{d}\mathbf{u}} + o(\sqrt{\rho}) \tag{A.23}$$

$$= o(\sqrt{\rho}) = \mathcal{O}(\sqrt{\rho}). \tag{A.24}$$

A.4 Proof of Theorem 1

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} = \frac{1}{2}\int_{\mathbb{R}^d} \left|\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta})\right| \mathrm{d}\boldsymbol{\theta} \tag{A.25}$$

$$= \frac{1}{2}\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left|\frac{Z_\pi}{Z_{\pi_\rho}}\mathcal{K}(\boldsymbol{\theta}) - 1\right| \mathrm{d}\boldsymbol{\theta}, \tag{A.26}$$

where Z_π and Z_{π_ρ} are the normalizing constants associated to π and π_ρ, respectively, and

$$\mathcal{K}(\boldsymbol{\theta}) = \frac{\pi_\rho(\boldsymbol{\theta}) Z_{\pi_\rho}}{\pi(\boldsymbol{\theta}) Z_\pi} \tag{A.27}$$

$$= \int_{\mathbb{R}^d} \exp\left(f(\boldsymbol{\theta}) - f(\mathbf{z}) - \frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathbf{z}\right\|^2\right) \mathrm{d}\mathbf{z}. \tag{A.28}$$

Note that

$$\int_{\mathbb{R}^d} \mathcal{K}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \frac{Z_{\pi_\rho}}{Z_\pi}. \tag{A.29}$$

Since f is assumed to be L-Lipschitz on $\mathbb{R}^d$, we have

$$\mathcal{K}(\boldsymbol{\theta}) \leq \int_{\mathbb{R}^d} \exp\left(L\left\|\boldsymbol{\theta} - \mathbf{z}\right\| - \frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathbf{z}\right\|^2\right) \mathrm{d}\mathbf{z}. \tag{A.30}$$

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

$$\mathcal{K}(\boldsymbol{\theta}) \leq \int_{\mathbb{R}^d} \exp\left(L\left\|\mathbf{u}\right\| - \frac{1}{2\rho^2}\left\|\mathbf{u}\right\|^2\right) \mathrm{d}\mathbf{u}. \tag{A.31}$$

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

$$\mathcal{K}(\boldsymbol{\theta}) \leq \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty t^{d-1} \exp\left(Lt - \frac{1}{2\rho^2}t^2\right) \mathrm{d}t. \tag{A.32}$$

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

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz - x^2/2} x^{d-1} \mathrm{d}x. \tag{A.33}$$

Then,

$$\mathcal{K}(\boldsymbol{\theta}) \leq A(\rho), \tag{A.34}$$

where

$$A(\rho) = \frac{2\pi^{d/2}\rho^d \Gamma(d) \exp\left(\frac{L^2\rho^2}{4}\right)}{\Gamma\left(\frac{d}{2}\right)} D_{-d}\left(-L\rho\right). \tag{A.35}$$

Then, with (A.29) and (A.34), we also have

$$\frac{Z_\pi}{Z_{\pi_\rho}} \geq \frac{1}{A(\rho)}. \tag{A.36}$$

We now use the triangle inequality in (A.26) which leads to

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \frac{1}{2}\left(\int_{\mathbb{R}^d}\left|\frac{Z_\pi}{Z_{\pi_\rho}}\mathcal{K}(\boldsymbol{\theta}) - \frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta})\right|\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \int_{\mathbb{R}^d}\left|\frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta}) - 1\right|\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right) \tag{A.37}$$

$$= \frac{1}{2}\left(\int_{\mathbb{R}^d}\left(\frac{Z_\pi}{Z_{\pi_\rho}}\mathcal{K}(\boldsymbol{\theta}) - \frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta})\right)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \int_{\mathbb{R}^d}\left(1 - \frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta})\right)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right). \tag{A.38}$$

The first term in this upper bound writes

$$\int_{\mathbb{R}^d}\left(\frac{Z_\pi}{Z_{\pi_\rho}} - \frac{1}{A(\rho)}\right)\mathcal{K}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1 - \frac{1}{A(\rho)}\int_{\mathbb{R}^d}\mathcal{K}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{A.39}$$

$$= \int_{\mathbb{R}^d}\left(1 - \frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta})\right)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{A.40}$$

This allows us to bound (A.38), that is

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \int_{\mathbb{R}^d}\left(1 - \frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta})\right)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{A.41}$$

Using one more time the L-Lipschitz assumption on f, we have for all $\boldsymbol{\theta}, \mathbf{z} \in \mathbb{R}^d$,

$$-\left(f(\mathbf{z}) - f(\boldsymbol{\theta})\right) \geq -\left|f(\mathbf{z}) - f(\boldsymbol{\theta})\right| \geq -L\left\|\boldsymbol{\theta} - \mathbf{z}\right\|, \tag{A.42}$$

$$\text{so that } \mathcal{K}(\boldsymbol{\theta}) \geq \int_{\mathbb{R}^d}\exp\left(-L\left\|\boldsymbol{\theta} - \mathbf{z}\right\| - \frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathbf{z}\right\|^2\right)\mathrm{d}\mathbf{z}. \tag{A.43}$$

With the same changes of variables as above, it follows

$$\mathcal{K}(\boldsymbol{\theta}) \geq B(\rho), \tag{A.44}$$

where

$$B(\rho) = \frac{2\pi^{d/2}\rho^d\Gamma(d)\exp\left(\frac{L^2\rho^2}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}D_{-d}\left(L\rho\right). \tag{A.45}$$

Then we have $1 - \frac{1}{A(\rho)}\mathcal{K}(\boldsymbol{\theta}) \leq 1 - \frac{B(\rho)}{A(\rho)}$ which combined with (A.41)

yields

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq 1 - \frac{D_{-d}\left(L\rho\right)}{D_{-d}\left(-L\rho\right)}. \tag{A.46}$$

A.5 Proof of Corollary 2

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

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz-x^2/2} x^{d-1} \mathrm{d}x. \tag{A.47}$$

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

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-x^2/2} x^{d-1} (1 - xz + o(z)) \mathrm{d}x \tag{A.48}$$

$$= \frac{\exp(-z^2/4)}{\Gamma(d)} \left(\int_0^{+\infty} e^{-x^2/2} x^{d-1} \mathrm{d}x - z \int_0^{+\infty} e^{-x^2/2} x^{d} \mathrm{d}x + o(z) \right) \tag{A.49}$$

$$= \frac{\exp(-z^2/4)}{\Gamma(d)} \left(\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} - z\Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(z) \right), \tag{A.50}$$

recording that $\int_0^{+\infty} e^{-x^2/2} x^d \mathrm{d}x = \Gamma((d+1)/2) 2^{d/2-1/2}$ (Gradshteyn and Ryzhik, 2015, Formula 3.383 11.). Using (A.50) for $z = \pm\rho L$ yields

$$1 - \frac{D_{-d}(L\rho)}{D_{-d}(-L\rho)} = 1 - \frac{\frac{\exp(-(\rho L)^2/4)}{\Gamma(d)} \left(\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} - \rho L \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(\rho) \right)}{\frac{\exp(-(\rho L)^2/4)}{\Gamma(d)} \left(\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} + \rho L \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(\rho) \right)} \tag{A.51}$$

$$= 1 - \frac{\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} - \rho L \Gamma\left(\frac{d+1}{2}\right) 2^{d/2-1/2} + o(\rho)}{\Gamma\left(\frac{d}{2}\right) 2^{d/2-1} \left(1 + \rho \frac{L\Gamma\left(\frac{d+1}{2}\right)\sqrt{2}}{\Gamma\left(\frac{d}{2}\right)} + o(\rho) \right)} \tag{A.52}$$

$$= 1 - \left(1 - \rho \frac{L\Gamma\left(\frac{d+1}{2}\right)\sqrt{2}}{\Gamma\left(\frac{d}{2}\right)} + o(\rho) \right) \left(1 - \rho \frac{L\Gamma\left(\frac{d+1}{2}\right)\sqrt{2}}{\Gamma\left(\frac{d}{2}\right)} + o(\rho) \right) \tag{A.53}$$

$$= \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L\rho + o(\rho). \tag{A.54}$$

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

$$\Gamma(z+1/2) \underset{z\to+\infty}{\sim} \sqrt{2\pi} z^z e^{-z} \tag{A.55}$$

$$\Gamma(z) \underset{z\to+\infty}{\sim} \sqrt{2\pi} z^{z-1/2} e^{-z}. \tag{A.56}$$

So that when d is large

$$\frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L\rho \underset{d\to+\infty}{\sim} \frac{2\sqrt{2}\sqrt{2\pi}(d/2)^{d/2} e^{-d/2}}{\sqrt{2\pi}(d/2)^{d/2-1/2} e^{-d/2}} L\rho \tag{A.57}$$

$$\underset{d\to+\infty}{\sim} 2\sqrt{2}(d/2)^{1/2} L\rho \tag{A.58}$$

$$\underset{d\to+\infty}{\sim} 2L\rho d^{1/2}. \tag{A.59}$$

A.6 Proof of Proposition 4

By using (A.34) and (A.44) we have for all $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$B(\rho) \leq \int_{\mathbb{R}^d} \exp\left(f(\boldsymbol{\theta}) - f(\mathbf{z}) - \frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathbf{z}\right\|^2\right) \mathrm{d}\mathbf{z} \quad \leq A(\rho) \tag{A.60}$$

$$B(\rho)\exp(-f(\boldsymbol{\theta})) \leq \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathbf{z}\right\|^2\right) \mathrm{d}\mathbf{z} \quad \leq A(\rho)\exp(-f(\boldsymbol{\theta})) \tag{A.61}$$

$$-\log A(\rho) + f(\boldsymbol{\theta}) \leq -\log \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathbf{z}\right\|^2\right) \mathrm{d}\mathbf{z} \quad \leq -\log B(\rho) + f(\boldsymbol{\theta}) \tag{A.62}$$

So that

$$-\log A(\rho) + \frac{d}{2}\log(2\pi\rho^2) \leq f_\rho(\boldsymbol{\theta}) - f(\boldsymbol{\theta}) \leq -\log B(\rho) + \frac{d}{2}\log(2\pi\rho^2). \tag{A.63}$$

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

A.7 Proof of Proposition 5

By using (A.34) and (A.44) it follows, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, that

$$B(\rho) \leq \mathcal{K}(\boldsymbol{\theta}) \leq A(\rho) \tag{A.64}$$

$$B(\rho)Z_\pi\pi(\boldsymbol{\theta}) \leq \mathcal{K}(\boldsymbol{\theta})Z_\pi\pi(\boldsymbol{\theta}) \leq A(\rho)Z_\pi\pi(\boldsymbol{\theta}). \tag{A.65}$$

Using (A.28) yields

$$B(\rho)\pi(\boldsymbol{\theta}) \leq \pi_\rho(\boldsymbol{\theta})\frac{Z_{\pi_\rho}}{Z_\pi} \leq A(\rho)\pi(\boldsymbol{\theta}) \tag{A.66}$$

$$B(\rho)\pi(\boldsymbol{\theta}) \leq \pi_\rho(\boldsymbol{\theta})(2\pi\rho^2)^{d/2} \leq A(\rho)\pi(\boldsymbol{\theta}). \tag{A.67}$$

Using (A.35) and (A.45) gives

$$\frac{N_\rho}{D_{-d}(-L\rho)}\pi_\rho(\boldsymbol{\theta}) \leq \pi(\boldsymbol{\theta}) \leq \frac{N_\rho}{D_{-d}(L\rho)}\pi_\rho(\boldsymbol{\theta}), \tag{A.68}$$

where

$$N_\rho = \frac{2^{d/2-1}\Gamma\left(d/2\right)}{\Gamma(d)\exp\left(L^2\rho^2/4\right)}.$$

Let $\mathcal{C}_\alpha^\rho$ an arbitrary $(1-\alpha)$-credibility region under π_ρ. By integrating (A.68) on $\mathcal{C}_\alpha^\rho$,

$$\frac{N_\rho}{D_{-d}(-L\rho)}(1-\alpha) \leq \int_{\mathcal{C}_\alpha^\rho} \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq \frac{N_\rho}{D_{-d}(L\rho)}(1-\alpha). \tag{A.69}$$

Since $\mathcal{C}_\alpha^\rho \subseteq \mathbb{R}^d$ and $\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$, the upper bound in (A.69) can be

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

A.8 Proof of Theorem 2

By using the notations $f(\boldsymbol{\theta})_- = -\min(f(\boldsymbol{\theta}), 0)$ and $f(\boldsymbol{\theta})_+ = \max(f(\boldsymbol{\theta}), 0)$, note that

$$\begin{aligned}\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} &= \frac{1}{2}\int_{\boldsymbol{\theta}\in\mathbb{R}^d} |\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})| \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_+ \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(1 - \frac{\pi_\rho(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)_+ \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{A.70}$$

since

$$\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_+ \mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} = 0,$$
$$|\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})| = (\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta}))_+ + (\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta}))_-.$$

Let

$$f_\rho(\boldsymbol{\theta}) := -\log \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right) \cdot \frac{\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}}. \tag{A.71}$$

The corresponding target distribution is thus

$$\pi_\rho(\boldsymbol{\theta}) = \frac{\exp(-f_\rho(\boldsymbol{\theta}))}{Z_{\pi_\rho}}, \text{ for a normalizing constant } Z_{\pi_\rho} = \int_{\mathbb{R}^d} \exp(-f_\rho(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}.$$

Similarly, we have

$$\pi(\boldsymbol{\theta}) = \frac{\exp(-f(\boldsymbol{\theta}))}{Z_\pi}, \text{ for a normalizing constant } Z_\pi = \int_{\mathbb{R}^d} \exp(-f(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}.$$

By substituting these into the bound (A.70), we have

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} = \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(1 - \exp\left(f(\boldsymbol{\theta}) - f_\rho(\boldsymbol{\theta})\right) \cdot \frac{Z_\pi}{Z_{\pi_\rho}}\right)_+ \mathrm{d}\boldsymbol{\theta} \tag{A.72}$$

using the fact that $(1-\exp(x))_+ \leq x_-$ for any $x \in \mathbb{R}$,

$$\leq \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\log\left(\frac{Z_\pi}{Z_{\pi_\rho}}\right) + \left(f(\boldsymbol{\theta}) - f_\rho(\boldsymbol{\theta})\right)\right)_- \mathrm{d}\boldsymbol{\theta}. \tag{A.73}$$

From (A.71), it is clear that

$$\exp(f(\boldsymbol{\theta}) - f_\rho(\boldsymbol{\theta})) = \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(f(\boldsymbol{\theta}) - f(\mathbf{z}) - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right) \cdot \frac{\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}}. \tag{A.74}$$

We now use (A_2), that is the convexity of f, which yields

$$f(\boldsymbol{\theta}) - f(\mathbf{z}) \leq \nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta} - \mathbf{z}). \tag{A.75}$$

Then, it follows that

$$\begin{aligned}
&\exp\left(f(\boldsymbol{\theta}) - f_\rho(\boldsymbol{\theta})\right) \\
&\leq (2\pi\rho^2)^{-d/2} \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(\nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta}-\mathbf{z}) - \frac{1}{2\rho^2}\|\boldsymbol{\theta}-\mathbf{z}\|^2\right) \mathrm{d}\mathbf{z} \\
&= \exp\left(\frac{\rho^2}{2}\left\|\nabla f(\boldsymbol{\theta})\right\|^2\right) \coloneqq \exp(\overline{B}(\boldsymbol{\theta})).
\end{aligned} \tag{A.76}$$

Using (A_1), and descent lemma, it follows that

$$f(\boldsymbol{\theta}) - f(\mathbf{z}) \geq \nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta}-\mathbf{z}) - \frac{M}{2}\|\boldsymbol{\theta}-\mathbf{z}\|^2. \tag{A.77}$$

Hence, using (A.74), we have

$$\begin{aligned}
&\exp\left(f(\boldsymbol{\theta}) - f_\rho(\boldsymbol{\theta})\right) \\
&\geq (2\pi\rho^2)^{-d/2} \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(\nabla f(\boldsymbol{\theta})^T(\boldsymbol{\theta}-\mathbf{z}) - \left(\frac{1+\rho^2 M}{2\rho^2}\right)\|\boldsymbol{\theta}-\mathbf{z}\|^2\right) \mathrm{d}\mathbf{z} \\
&= \left(\exp\left(\frac{\rho^2}{2(1+\rho^2 M)}\left\|\nabla f(\boldsymbol{\theta})\right\|^2\right)\left(\frac{1}{1+\rho^2 M}\right)^{d/2}\right) \qquad \text{(A.78)} \\
&= \exp\left(\frac{\rho^2}{2(1+\rho^2 M)}\left\|\nabla f(\boldsymbol{\theta})\right\|^2 - \frac{d}{2}\log(1+\rho^2 M)\right) \coloneqq \exp(\underline{B}(\boldsymbol{\theta})).
\end{aligned} \tag{A.79}$$

One can also show that we have

$$\begin{aligned}
&\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp(-f_\rho(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}\in\mathbb{R}^d}\int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right) \cdot \frac{\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}}\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp(-f(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

Hence, in this case $Z_\pi = Z_{\pi_\rho}$. By combining this and (A.79) with our

bound (A.73), we obtain

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\underline{B}(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} \tag{A.80}$$

$$= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\frac{\rho^2 \left\|\nabla U_1(\mathbf{A}_1\boldsymbol{\theta})\right\|^2}{2(1+\rho^2 M_1)} - \frac{d}{2}\log(1+\rho^2 M_1) \right)_- \mathrm{d}\boldsymbol{\theta} \tag{A.81}$$

$$\leq \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(-\frac{d}{2}\log(1+\rho^2 M_1) \right)_- \mathrm{d}\boldsymbol{\theta} \tag{A.82}$$

$$\leq \frac{d}{2} M_1 \rho^2. \tag{A.83}$$

A.9 Proof of Corollary 3

Equation (A.28) becomes

$$\mathcal{K}(\boldsymbol{\theta}) = \prod_{i=1}^{b} \int_{\mathbb{R}^{d_i}} \exp\left(f_i(\mathbf{A}_i\boldsymbol{\theta}) - f_i(\mathbf{z}_i) - \frac{1}{2\rho^2}\left\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\right\|^2 \right) \mathrm{d}\mathbf{z}_i = \prod_{i=1}^{b} \mathcal{K}_i(\boldsymbol{\theta}). \tag{A.84}$$

Bounding each term in (A.84) and following the proof of Theorem 1 detailed in Appendix A.4 completes the proof.

A.10 Proof of Proposition 7

Assume without loss of generality that $\pi(\boldsymbol{\theta})$ is normalized, i.e., $\int_{\mathbb{R}^d} \exp(-f(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta} = 1$ (if it is not, we can fix it by adding the logarithm of the normalizing constant). Then the distribution

$$\pi_\rho(\boldsymbol{\theta}) = \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{\left\|\boldsymbol{\theta}-\mathbf{z}\right\|^2}{2\rho^2} \right) \mathrm{d}\mathbf{z} \tag{A.85}$$

is the convolution of $\pi(\boldsymbol{\theta}) = \exp(-f(\boldsymbol{\theta}))$ and a d-dimensional Gaussian random variable with mean zero and covariance $\rho^2\mathbf{I}_d$. In particular, it is clear that

$$\int_{\mathbb{R}^d} \pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_{\mathbb{R}^d} \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-f(\mathbf{z}) - \frac{\left\|\boldsymbol{\theta}-\mathbf{z}\right\|^2}{2\rho^2} \right) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{z} \tag{A.86}$$

$$= \int_{\mathbb{R}^d} \exp(-f(\mathbf{z})) = 1. \tag{A.87}$$

The first part of the bound follows from the fact that the expectation of the norm of this Gaussian random variable is bounded by $\rho\sqrt{d}$. One can also retrieve this result by applying Proposition 2 to the Gaussian smoothing kernel.

In order to obtain the second part, we are going to use the dual formulation of the 1-Wasserstein distance, see e.g., Remark 6.5 of Villani

(2008). We have:

$$W_1(\pi, \pi_\rho) = \sup_{g:\|g\|_{\text{Lip}} \leq 1} \int_{\mathbb{R}^d} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \int_{\mathbb{R}^d} g(\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$
$$= \sup_{g\in\mathcal{C}^1(\mathbb{R}^d):\|\nabla g\|_\infty \leq 1} \int_{\mathbb{R}^d} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \int_{\mathbb{R}^d} g(\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{A.88}$$

where the second equality follows from the fact that differentiable functions g with $\|\nabla g\|_\infty \leq 1$ are dense among 1-Lipschitz functions on $\mathbb{R}^d$. The evolution of a density π_ρ as we increase the variance ρ^2 is known to follow the heat equation (Lawler, 2010, Section 2.4), that is

$$\frac{\mathrm{d}}{\mathrm{d}(\rho^2)}\pi_\rho(\boldsymbol{\theta}) = \frac{1}{2}\Delta\pi_\rho(\boldsymbol{\theta}), \tag{A.89}$$

where $\Delta\pi_\rho(\boldsymbol{\theta}) = \sum_{i=1}^d \frac{\partial^2}{\partial\theta_i^2}\pi_\rho(\boldsymbol{\theta})$ denotes the Laplacian of π_ρ. By integration, we obtain:

$$\sup_{g\in\mathcal{C}^1(\mathbb{R}^d):\|\nabla g\|_\infty < 1} \frac{\mathrm{d}}{\mathrm{d}(\rho^2)} \int_{\mathbb{R}^d} g(\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \sup_{g\in\mathcal{C}^1(\mathbb{R}^d):\|\nabla g\|_\infty \leq 1} \frac{1}{2}\int_{\mathbb{R}^d} g(\boldsymbol{\theta})\Delta\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{A.90}$$

Now if we define the functional

$$\mathcal{F}(\mu) := \sup_{g\in\mathcal{C}^1(\mathbb{R}^d):\|\nabla g\|_\infty \leq 1} \frac{1}{2}\int_{\mathbb{R}^d} g(\boldsymbol{\theta})\Delta\mu(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{A.91}$$

Then it is easy to see that this functional is convex (i.e., $\mathcal{F}((\mu+\nu)/2) \leq \frac{\mathcal{F}(\mu)+\mathcal{F}(\nu)}{2}$) and shift-invariant (i.e., if $\nu(\mathbf{x}) = \mu(\mathbf{x}-\mathbf{a})$, with some constant vector $\mathbf{a} \in \mathbb{R}^d$, then $\mathcal{F}(\nu) = \mathcal{F}(\mu)$). Therefore, it follows by the argument on pages 1-2 of Bennett and Bez (2015) (monotonicity property of the heat semigroup for convex functionals) that $\mathcal{F}(\pi_\rho) \leq \mathcal{F}(\pi)$ for every $\rho \geq 0$.

Initially, we have

$$\mathcal{F}(\pi) = \sup_{g\in C^1(\mathbb{R}^d):\|\nabla g\|_\infty \leq 1} \frac{1}{2}\sum_{i=1}^d \int_{\mathbb{R}^d} g(\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i^2}\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{A.92}$$

After separating $\boldsymbol{\theta}$ to $\theta_i \in \mathbb{R}$ and $\boldsymbol{\theta}_{-i} \in \mathbb{R}^{d-1}$ (denoting the rest of the coordinates), we have:

$$\int_{\mathbb{R}^d} g(\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i^2}\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} g(\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i^2}\pi(\boldsymbol{\theta})\mathrm{d}\theta_i\right]\mathrm{d}\boldsymbol{\theta}_{-i}. \tag{A.93}$$

Now, integrating by parts and using the fact that f satisfies

$$f(\boldsymbol{\theta}) \geq a_1 + a_2\|\boldsymbol{\theta}\|^\alpha \quad \text{and} \quad \|\nabla f(\boldsymbol{\theta})\| \leq M\|\boldsymbol{\theta}\|, \tag{A.94}$$

for some $a_1 \in \mathbb{R}$, $a_2 > 0$, $\alpha > 0$, along with the Lipschitz continuity of g

leads to

$$\int_{\mathbb{R}} g(\boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i^2} \pi(\boldsymbol{\theta}) \mathrm{d}\theta_i = \left[-g(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \exp(-f(\boldsymbol{\theta})) \right]_{\theta_i=-\infty}^{\theta_i=\infty} + \int_{\mathbb{R}} \frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \exp(-f(\boldsymbol{\theta})) \mathrm{d}\theta_i \tag{A.95}$$

$$= \int_{\mathbb{R}} \frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\theta_i. \tag{A.96}$$

By summing up for $i = 1$ to $i = d$, we obtain:

$$\mathcal{F}(\pi) \leq \frac{1}{2} \sup_{g \in \mathcal{C}^1(\mathbb{R}^d): \|\nabla g\|_\infty \leq 1} \sum_{i=1}^{d} \int_{\mathbb{R}^d} \frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{A.97}$$

$$\leq \frac{1}{2} \int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{\theta})\| \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{A.98}$$

Using the monotonicity property of $\mathcal{F}(\pi_\rho)$, now the second bound of the theorem follows based on formula (A.88).

For the integral of the norm of the gradient, we have by Jensen's inequality that

$$\int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{\theta})\| \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq \left(\int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{\theta})\|^2 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right)^{1/2}. \tag{A.99}$$

For some index $1 \leq i \leq d$, we have:

$$\int_{\mathbb{R}^d} \left(\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \right)^2 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \right)^2 \exp\left(-f(\boldsymbol{\theta})\right) \mathrm{d}\theta_i \right] \mathrm{d}\boldsymbol{\theta}_{-i}, \tag{A.100}$$

and using integration by parts, the growth, smoothness and convexity conditions we have:

$$\int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \right)^2 \exp\left((-f(\boldsymbol{\theta})) \right) \mathrm{d}\theta_i \tag{A.101}$$

$$= \left[-\exp(-f(\boldsymbol{\theta})) \frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) \right]_{\theta_i=-\infty}^{\theta_i=\infty} + \int_{\mathbb{R}} \exp(-f(\boldsymbol{\theta})) \frac{\partial^2}{\partial \theta_i^2} f(\boldsymbol{\theta}) \mathrm{d}\theta_i \tag{A.102}$$

$$\leq \int_{\mathbb{R}} \exp(-f(\boldsymbol{\theta})) M \mathrm{d}\theta_i. \tag{A.103}$$

By integrating this out according to $\boldsymbol{\theta}_{-i}$ and summing up from $i = 1$ to $i = d$, we obtain that $\int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{\theta})\|^2 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq Md$, so the last claim of the theorem follows.

A.11 Proof of Proposition 8

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} = \frac{1}{2}\int_{\mathbb{R}^d} |\pi_\rho(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta})| \mathrm{d}\boldsymbol{\theta} \tag{A.104}$$

$$= \frac{1}{2\mathrm{Vol}(\mathcal{K})}\int_{\mathbb{R}^d} \left| e^{-\iota_{\mathcal{K}}(\boldsymbol{\theta})} - \frac{\displaystyle\int_{\mathcal{K}} \exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}} \right| \mathrm{d}\boldsymbol{\theta} \tag{A.105}$$

$$= \frac{1}{2\mathrm{Vol}(\mathcal{K})}\left[\int_{\mathcal{K}} \left| 1 - \frac{\displaystyle\int_{\mathcal{K}} \exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}} \right| \mathrm{d}\boldsymbol{\theta} + \int_{\mathcal{K}^c} \left| 0 - \frac{\displaystyle\int_{\mathcal{K}} \exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}} \right| \mathrm{d}\boldsymbol{\theta}\right] \tag{A.106}$$

$$= \frac{1}{\mathrm{Vol}(\mathcal{K})}\int_{\mathcal{K}^c} \frac{\displaystyle\int_{\mathcal{K}} \exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}} \mathrm{d}\boldsymbol{\theta} \tag{A.107}$$

$$= \frac{1}{\mathrm{Vol}(\mathcal{K})}\int_{\mathcal{K}^c} \left[\frac{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\iota_{\mathcal{K}}(\mathbf{z}) - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}}\right] \mathrm{d}\boldsymbol{\theta} \tag{A.108}$$

$$= \frac{1}{\mathrm{Vol}(\mathcal{K})}\int_{\mathcal{K}^c} \left[\exp\left(\log \frac{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\iota_{\mathcal{K}}(\mathbf{z}) - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}}{(2\pi\rho^2)^{d/2}}\right)\right] \mathrm{d}\boldsymbol{\theta} \tag{A.109}$$

$$\overset{(1)}{\leq} \frac{1}{\mathrm{Vol}(\mathcal{K})}\int_{\mathcal{K}^c} \exp\left(-\frac{1}{2\rho^2}\left\|\boldsymbol{\theta} - \mathrm{proj}_{\mathcal{K}}(\boldsymbol{\theta})\right\|^2\right) \mathrm{d}\boldsymbol{\theta} \tag{A.110}$$

$$\overset{(2)}{\leq} \sum_{i=1}^{d} \left(\frac{\sqrt{2}\rho d}{r}\right)^i. \tag{A.111}$$

To obtain (1), we used the convexity of $\iota_{\mathcal{K}}$ and Proposition 13 in Chapter 5. The last inequality (2) and (1.47) both follow from Brosse et al. (2017, Proof of Proposition 4.b)).

Appendices – Chapter 2

Chapter contents

B.1 Extended state space Langevin dynamics

When the functions $(f_i; i \in [b])$ are continuously differentiable, we point out here another possible approach to sample from the joint distribution π_ρ in (2.2) based on overdamped Langevin dynamics. The associated stochastic differential equation (SDE) writes

$$\mathrm{d}\begin{pmatrix} \boldsymbol{\theta}_t \\ \mathbf{z}_{1,t} \\ \vdots \\ \mathbf{z}_{p,t} \end{pmatrix} = -\begin{pmatrix} \sum_{i=p+1}^{b} \nabla f_i(\mathbf{A}_i\boldsymbol{\theta}_t) + \rho^{-2}\sum_{i=1}^{p} \mathbf{A}_i^T\left(\mathbf{A}_i\boldsymbol{\theta}_t - \mathbf{z}_{i,t}\right) \\ \rho^{-2}\left(\mathbf{z}_{1,t} - \mathbf{A}_1\boldsymbol{\theta}_t\right) + \nabla f_1(\mathbf{z}_{1,t}) \\ \vdots \\ \rho^{-2}\left(\mathbf{z}_{p,t} - \mathbf{A}_p\boldsymbol{\theta}_t\right) + \nabla f_p(\mathbf{z}_{p,t}) \end{pmatrix} \mathrm{d}t + \sqrt{2}\begin{pmatrix} \mathrm{d}\boldsymbol{\xi}_t \\ \mathrm{d}\boldsymbol{\xi}_{1,t} \\ \vdots \\ \mathrm{d}\boldsymbol{\xi}_{p,t} \end{pmatrix}, \tag{B.1}$$

where $(\boldsymbol{\xi}_t)_{t\geq 0}$ and $(\boldsymbol{\xi}_{i,t})_{t\geq 0}$ are independent d-dimensional and d_i-dimensional Brownian motions, respectively. By introducing the process $(\mathbf{s}_t)_{t\geq 0} = (\boldsymbol{\theta}_t, \mathbf{z}_{1,t}, \ldots, \mathbf{z}_{p,t})_{t\geq 0}$, the SDE (B.1) writes

$$\mathrm{d}\mathbf{s}_t = -\nabla U(\mathbf{s}_t) + \sqrt{2}\mathrm{d}\boldsymbol{\xi}_t^{'},$$

where

$$U(\mathbf{s}_t) = \sum_{i=p+1}^{b} f_i(\mathbf{A}_i\boldsymbol{\theta}_t) + \sum_{i=1}^{p} f_i(\mathbf{z}_{i,t}) + \frac{1}{2\rho^2}\left\|\mathbf{A}_i\boldsymbol{\theta}_t - \mathbf{z}_{i,t}\right\|^2,$$

and $(\boldsymbol{\xi}_t^{'})_{t\geq 0}$ is a $(d+k)$-dimensional Brownian motion, where $k = \sum_{i=1}^{p} d_i$.

Similarly to Algorithm 1, the SDE (B.1) leads to a divide-to-conquer implementation since each auxiliary variable $\mathbf{z}_{i,t}$ can be sampled independently from the others given the current iterate $\boldsymbol{\theta}_t$. An interesting advantage of working with (B.1) is that, contrary to SGS, the update of $\mathbf{s}_t$ can be undertaken in a fully parallel manner instead of a sequential one.

Appendices – Chapter 3

Chapter contents

C.1 Proof of Theorem 3

The following two propositions are going to be used for the proof of Theorem 3. The first one will allow us to bound the Wasserstein distance of two log-concave distributions based on the differences between their gradients. This is achieved by coupling processes evolving according to the Langevin dynamics with common Brownian noise.

Proposition 17. *Let μ and μ' be two distributions on $\mathbb{R}^n$ that are absolutely continuous with respect to the Lebesgue measure, and whose negative log-likelihoods are continuously differentiable, strongly convex and smooth (gradient-Lipschitz). Denote the strong convexity constants $m(\mu), m(\mu')$ and smoothness constants $M(\mu)$ and $M(\mu')$. Then the Wasserstein distance of order $1 \leq p \leq \infty$ of these two distributions can be upper bounded as*

$$W_p(\mu, \mu') \leq \frac{\|D_{\mu,\mu'}\|_{L^p(\mu)}}{m(\mu')} \quad \text{for} \quad D_{\mu,\mu'}(\mathbf{z}) = \nabla \log \mu(\mathbf{z}) - \nabla \log \mu'(\mathbf{z}). \tag{C.1}$$

Proof. Let $\mu(\mathbf{z}) = \exp(-U(\mathbf{z}))$ and $\mu'(\mathbf{z}) = \exp(-U'(\mathbf{z}))$.

First, we are going to consider the case $1 \leq p < \infty$. Note that it is easy to show that under the strong convexity and smoothness assumptions of this proposition, the Wasserstein distance of order p between μ and μ' is finite for such p. Assume that $(\mathbf{X}_1(0), \mathbf{X}_3(0))$ is an optimal coupling in Wasserstein distance of order p between μ and μ', so that $\mathbf{X}_1(0) \sim \mu$, $\mathbf{X}_3(0) \sim \mu'$, and

$$\left[\mathbb{E}\left(\|\mathbf{X}_1(0) - \mathbf{X}_3(0)\|^p\right)\right]^{1/p} = W_p(\mu, \mu'). \tag{C.2}$$

The existence of such a coupling follows from Theorem 4.1 by Villani (2008). Let $\mathbf{X}_2(0) = \mathbf{X}_1(0)$. We now define three Langevin diffusions

$(\mathbf{X}_1(t), \mathbf{X}_2(t), \mathbf{X}_3(t))_{t\geq 0}$ with a common noise (synchronous coupling)

$$\mathrm{d}\mathbf{X}_1(t) = -\nabla U(\mathbf{X}_1(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{C.3}$$

$$\mathrm{d}\mathbf{X}_2(t) = -\nabla U'(\mathbf{X}_2(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{C.4}$$

$$\mathrm{d}\mathbf{X}_3(t) = -\nabla U'(\mathbf{X}_3(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t. \tag{C.5}$$

Under the strong convexity and smoothness assumptions on the log-densities, these SDEs admit unique strong solutions (see Theorem 3.1 by Pavliotis (2014) and Arnold (1974)). Since $\mathbf{X}_1(0) \sim \mu$ and $\mathbf{X}_3(0) \sim \mu'$, we can see that $\mathbf{X}_1(t) \sim \mu$ and $\mathbf{X}_3(t) \sim \mu'$ for every $t \geq 0$. $\mathbf{X}_2(t)$ is initialized at μ since $\mathbf{X}_2(0) = \mathbf{X}_1(0)$ and converges towards μ'. The proof of this proposition is based on a coupling argument based on these three diffusions. Let

$$D_{12}(t) = \mathbf{X}_1(t) - \mathbf{X}_2(t) - t(\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))). \tag{C.6}$$

Then we can decompose $\mathbf{X}_1(t) - \mathbf{X}_3(t)$ as

$$\mathbf{X}_1(t) - \mathbf{X}_3(t) = t(\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))) + D_{12}(t) + (\mathbf{X}_2(t) - \mathbf{X}_3(t)). \tag{C.7}$$

In the next two paragraphs of the proof, we are going to establish the following auxiliary inequalities

$$\|\mathbf{X}_2(t) - \mathbf{X}_3(t)\| \leq \exp(-m(\mu')t) \cdot \|\mathbf{X}_1(0) - \mathbf{X}_3(0)\|, \tag{C.8}$$

$$\begin{aligned} \|D_{12}(t)\| \; &\leq C_0 t^3 + C_1 t^2 (\|\nabla U(\mathbf{X}_1(0))\| + \|\nabla U'(\mathbf{X}_1(0))\|) \\ &\quad + C_2 t \sup_{0\leq s\leq t} \|\mathbf{B}_s\| \text{ for } 0 \leq t \leq C_3, \end{aligned} \tag{C.9}$$

for positive constants C_0, C_1, C_2, C_3 that only depend on the dimension d and the convexity parameters $m(\mu), m(\mu'), M(\mu), M(\mu')$. Let $\|X\|_{L^p} = (\mathbb{E}(\|X\|^p))^{1/p}$ denote the L^p norm of a random variable. By taking the L^p norms of both sides of (C.7), and using Minkowski's inequality, we can see that

$$\begin{aligned} \|\mathbf{X}_1(t) - \mathbf{X}_3(t)\|_{L^p} \leq t\|\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))\|_{L^p} + \|D_{12}(t)\|_{L^p} \\ + \|\mathbf{X}_2(t) - \mathbf{X}_3(t)\|_{L^p}. \end{aligned} \tag{C.10}$$

By the definition of the Wasserstein distance, we know that $W_p(\mu, \mu') \leq \|\mathbf{X}_1(t) - \mathbf{X}_3(t)\|_{L^p}$, and by assuming inequalities (C.8) and (C.9) are true,

we obtain that for $0 \leq t \leq C_3$,

$$W_p(\mu, \mu') \leq t\|\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))\|_{L^p} + W_p(\mu, \mu')\exp(-m(\mu')t) + C_0t^3 + C_1t^2(\|\nabla U(\mathbf{X}_1(0))\|_{L^p} + \|\nabla U'(\mathbf{X}_1(0))\|_{L^p}) + C_2t\left\|\sup_{0\leq s\leq t}\|\mathbf{B}_s\|\right\|_{L^p}. \tag{C.11}$$

It is easy to show that under the strong convexity and smoothness assumptions of this proposition, the terms $\|\nabla U(\mathbf{X}_1(0))\|_{L^p}$ and $\|\nabla U'(\mathbf{X}_1(0))\|_{L^p}$ are finite. By the reflection principle for the Brownian motion (see Lévy (1940)), in one dimension, the distribution of $\sup_{0\leq s\leq t} B_s$ is the same as the distribution of $|B_t|$. Using the triangle inequality, and the fact that $\|Y\|_{L^p} \leq \sqrt{p}$ for a standard Gaussian random variable Y, it follows that

$$\left\|\sup_{0\leq s\leq t}\|\mathbf{B}_s\|\right\|_{L^p} \leq 2d\sqrt{t}\sqrt{p}. \tag{C.12}$$

Hence all of the terms bounding $\|D_{12}(t)\|_{L^p}$ in (C.9) are of order $o(t)$, and the claim of the proposition follows by rearrangement and letting $t \searrow 0$.

Now we are going to prove the two auxiliary inequalities. We start with (C.8). From Itô's formula (see Lemma 3.2 by Pavliotis (2014)), $\|\mathbf{X}_2(t) - \mathbf{X}_3(t)\|^2$ is differentiable in t and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{X}_2(t) - \mathbf{X}_3(t)\|^2 = -2\left\langle \mathbf{X}_2(t) - \mathbf{X}_3(t), \nabla U'(\mathbf{X}_2(t) - \nabla U'(\mathbf{X}_3(t))\right\rangle \tag{C.13}$$

$$\leq -2m(\mu')\|\mathbf{X}_2(t) - \mathbf{X}_3(t)\|^2, \tag{C.14}$$

where the last step follows from the strong convexity of U'. We obtain (C.8) by Grönwall's inequality and rearrangement.

We continue with the proof of (C.9). By Itô's formula, we can see that

$$D_{12}(t) = \mathbf{X}_1(t) - \mathbf{X}_2(t) - t(\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))) \tag{C.15}$$

$$= \int_{s=0}^{t}(\nabla U'(\mathbf{X}_2(t)) - \nabla U(\mathbf{X}_1(t)))\mathrm{d}s - t(\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))) \tag{C.16}$$

$$= \int_{s=0}^{t}[\nabla U'(\mathbf{X}_2(t)) - \nabla U'(\mathbf{X}_1(0))]\mathrm{d}s + \int_{s=0}^{t}[\nabla U(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(t))]\mathrm{d}s. \tag{C.17}$$

Using the smoothness assumption for U and U', and the fact that $\mathbf{X}_1(0) =$

$\mathbf{X}_2(0)$, we have

$$\|D_{12}(t)\| \leq M(\mu') \int_0^t \|\mathbf{X}_2(t) - \mathbf{X}_2(0)\| \mathrm{d}s + M(\mu) \int_0^t \|\mathbf{X}_1(t) - \mathbf{X}_1(0)\| \mathrm{d}s. \tag{C.18}$$

Let

$$\mathbf{Y}_1'(t) = -\nabla U(\mathbf{Y}_1(t)), \tag{C.19}$$

$$\mathbf{Y}_2'(t) = -\nabla U'(\mathbf{Y}_2(t)), \tag{C.20}$$

and assume that $\mathbf{Y}_1(0) = \mathbf{Y}_2(0) = \mathbf{X}_1(0) = \mathbf{X}_2(0)$. Then these ODEs have a unique solution (see page 74 of Perko (2013)). Now by the triangle inequality, and the fact that $\mathbf{Y}_1(0) = \mathbf{X}_1(0)$, we have

$$\|\mathbf{X}_1(s) - \mathbf{X}_1(0)\| \leq \|\mathbf{Y}_1(s) - \mathbf{Y}_1(0)\| + \|\mathbf{Y}_1(s) - \mathbf{X}_1(s)\|. \tag{C.21}$$

For the first part, by Taylor's expansion, and the smoothness assumption on U, we have

$$\|\mathbf{Y}_1(t) - \mathbf{Y}_1(0)\| \leq s\|\nabla U(\mathbf{Y}_1(0))\| + \frac{1}{2} M(\mu) s^2. \tag{C.22}$$

For the second part, by Itô's formula, we have

$$\mathbf{Y}_1(s) - \mathbf{X}_1(s) = \int_{r=0}^{s} [\nabla U(\mathbf{X}_1(r)) - \nabla U(\mathbf{Y}_1(r))] \mathrm{d}r + \sqrt{2}\mathbf{B}_s, \tag{C.23}$$

$$\|\mathbf{Y}_1(s) - \mathbf{X}_1(s)\| \leq M(\mu) \int_{r=0}^{s} \|\mathbf{X}_1(r) - \mathbf{Y}_1(r)\| \mathrm{d}r + \sqrt{2}\|\mathbf{B}_s\| \tag{C.24}$$

$$\sup_{0 \leq r \leq s} \|\mathbf{Y}_1(s) - \mathbf{X}_1(s)\| \leq M(\mu) s \sup_{0 \leq r \leq s} \|\mathbf{Y}_1(s) - \mathbf{X}_1(s)\| + \sqrt{2} \sup_{0 \leq r \leq s} \|\mathbf{B}_r\|. \tag{C.25}$$

Hence for $s \leq 1/(2M(\mu))$, we have

$$\sup_{0 \leq r \leq s} \|\mathbf{Y}_1(s) - \mathbf{X}_1(s)\| \leq 2\sqrt{2} \sup_{0 \leq r \leq s} \|\mathbf{B}_r\|. \tag{C.26}$$

By combining the above two bounds, for $0 \leq s \leq 1/(2M(\mu))$, we have

$$\|\mathbf{X}_1(s) - \mathbf{X}_1(0)\| \leq s\|\nabla U(\mathbf{Y}_1(0))\| + \frac{1}{2} M(\mu) s^2 + 2\sqrt{2} \sup_{0 \leq r \leq s} \|\mathbf{B}_r\|, \tag{C.27}$$

and by the same argument, for $0 \le s \le 1/(2M(\mu'))$,

$$\|\mathbf{X}_2(s) - \mathbf{X}_2(0)\| \le s\|\nabla U'(W_2(0))\| + \frac{1}{2}M(\mu')s^2 + 2\sqrt{2}\sup_{0\le r\le s}\|\mathbf{B}_r\|. \tag{C.28}$$

Inequality now (C.9) follows by substituting these into (C.18) and doing some rearrangement.

Finally, the result for $p = \infty$ follows from a limiting argument. By Proposition 3 of Givens and Shortt (1984), we have

$$W_\infty(\mu,\mu') = \lim_{p\to\infty} W_p(\mu,\mu') \le \sup_{1\le p<\infty}\frac{\|D_{\mu,\mu'}\|_{L^p(\mu)}}{m(\mu')} \le \frac{\|D_{\mu,\mu'}\|_{L^\infty(\mu)}}{m(\mu')}. \tag{C.29}$$

□

Proposition 18. *Let $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ be two parameter values, and μ_i, resp. μ_i', denotes the conditional distributions of $\mathbf{z}_i$ given $\boldsymbol{\theta}$ under π_ρ, resp. $\boldsymbol{\theta}'$. Then under Assumption (A_3), for every $1 \le p \le \infty$, we have*

$$W_p(\mu_i,\mu_i') \le \frac{1}{1+\rho^2 m_i}\|\mathbf{A}_i(\boldsymbol{\theta}-\boldsymbol{\theta}')\|. \tag{C.30}$$

Proof. We have

$$\mu_i(\mathbf{z}) \propto \exp\left(-f_i(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}\|^2}{2\rho^2}\right) \tag{C.31}$$

$$\mu_i'(\mathbf{z}) \propto \exp\left(-f_i(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta}' - \mathbf{z}\|^2}{2\rho^2}\right). \tag{C.32}$$

Proposition 17 requires the smoothness (gradient Lipschitz) property, so it cannot be applied directly to these potentials under our assumptions. To overcome this difficulty, we are going to use the Moreau-Yosida envelope of f_i (see e.g., Durmus, Moulines, and Pereyra (2018)), defined for any regularization parameter $\lambda > 0$ as

$$f_i^\lambda(\mathbf{z}) := \min_{\mathbf{y}\in\mathbb{R}^d}\left\{f_i(\mathbf{y}) + (2\lambda)^{-1}\|\mathbf{y}-\mathbf{z}\|^2\right\}. \tag{C.33}$$

By Theorem 1.25 of Rockafellar and Wets (1998), f_i^λ converges pointwise to f_i, i.e., for every $\mathbf{z} \in \mathbb{R}^d$,

$$\lim_{\lambda\to 0} f_i^\lambda(\mathbf{z}) = f_i(\mathbf{z}). \tag{C.34}$$

Moreover, from Proposition 12.19 of Rockafellar and Wets (1998) and Theorem 2.2 of Lemaréchal and Sagastizábal (1997) it follows that f_i^λ is

λ^{-1} gradient Lipschitz and $\frac{m_i}{1+\lambda m_i}$-strongly convex. Let

$$\mu_i^\lambda(\mathbf{z}) \propto \exp\left(-f_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}\|^2}{2\rho^2}\right) \tag{C.35}$$

$$\mu_i'^\lambda(\mathbf{z}) \propto \exp\left(-f_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta}' - \mathbf{z}\|^2}{2\rho^2}\right), \tag{C.36}$$

then we have

$$\|\nabla \log(\mu_i^\lambda(\mathbf{z})) - \nabla \log(\mu_i'^\lambda(\mathbf{z}))\| = \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{A}_i\boldsymbol{\theta}'\|}{\rho^2}. \tag{C.37}$$

Since $-\log \mu_i^\lambda(\mathbf{z})$ and $-\log \mu_i'^\lambda(\mathbf{z})$ are $\frac{m_i}{1+\lambda m_i} + \frac{1}{\rho^2}$-strongly convex and $\frac{1}{\lambda} + \frac{1}{\rho^2}$-smooth, it follows from Proposition 17 that we have for every $1 \leq p \leq \infty$

$$W_p(\mu_i^\lambda, \mu_i'^\lambda) \leq \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{A}_i\boldsymbol{\theta}'\|}{1 + \rho^2 m_i/(1+m_i\lambda)}. \tag{C.38}$$

Now we are going to consider the case $1 \leq p < \infty$ first. To complete the proof, we still need to bound $W_p(\mu_i^\lambda, \mu_i)$. By Theorem 6.15 of Villani (2008), we have

$$W_p(\mu_i^\lambda, \mu_i) \leq \left[\int_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z} - \boldsymbol{\theta}\|^p |\mu_i(\mathbf{z}) - \mu_i^\lambda(\mathbf{z})| \mathrm{d}\mathbf{z}\right]^{1/p}. \tag{C.39}$$

Note that $|\mu_i(\mathbf{z}) - \mu_i^\lambda(\mathbf{z})| \leq \mu_i(\mathbf{z}) + \mu_i^\lambda(\mathbf{z})$. Moreover, from the definition of the Moreau-Yosida envelope f_i^λ, it follows that $f_i^\lambda(\mathbf{z}) \leq f_i^{\lambda'}(\mathbf{z})$ for $\lambda' < \lambda$, hence it is monotone increasing towards $f_i(\mathbf{z})$ as $\lambda \to 0$. This implies that the normalising constant

$$Z_i^\lambda = \int_{\mathbf{z}} \exp\left(-f_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{z} - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right) \mathrm{d}\mathbf{z} \tag{C.40}$$

is monotone decreasing towards $Z_i = \int_{\mathbf{z}} \exp\left(-f_i(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right) \mathrm{d}\mathbf{z}$ as $\lambda \to 0$ by the monotone convergence theorem. Therefore we have for any fixed $\Lambda > 0$ and $0 < \lambda < \Lambda$

$$\mu_i^\lambda(\mathbf{z}) = \frac{\exp\left(-f_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i^\lambda} \leq \frac{\exp\left(-f_i^\Lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i}. \tag{C.41}$$

This means that for $\lambda < \Lambda$, we have

$$\|\mathbf{z}-\boldsymbol{\theta}\|^p |\mu_i(\mathbf{z}) - \mu_i^\lambda(\mathbf{z})| \leq \|\mathbf{z}-\boldsymbol{\theta}\|^p \left(\mu_i(\mathbf{z}) + \frac{\exp\left(-f_i^\Lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i}\right). \tag{C.42}$$

Using the strong-convexity of $-\log \mu_i$, it follows that it has a unique minimizer which we denote by $\mathbf{z}_i^*$. In particular, we have

$$\int_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z}-\boldsymbol{\theta}\|^p \mu_i(\mathbf{z})\mathrm{d}\mathbf{z} \leq \mu_i(\mathbf{z}_i^*) \int_{z\in\mathbb{R}^d} \|\mathbf{z}-\boldsymbol{\theta}\|^p \exp\left(-(m_i+1/\rho^2)\|\mathbf{z}-\mathbf{z}_i^*\|^2/2\right)\mathrm{d}\mathbf{z} \tag{C.43}$$

$$< \infty, \tag{C.44}$$

and with the same argument we can also show that

$$\int_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z}-\boldsymbol{\theta}\|^p \frac{\exp\left(-f_i^\Lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i} < \infty. \tag{C.45}$$

Hence using the pointwise convergence (C.34) it follows from the dominated convergence theorem and the bound (C.39) that $W_p(\mu_i^\lambda, \mu_i) \to 0$ as $\lambda \to 0$. The same also holds for $W_p(\mu_i'^\lambda, \mu_i')$, so we can conclude using (C.38) and the triangle inequality

$$W_p(\mu_i, \mu_i') \leq W_p(\mu_i, \mu_i^\lambda) + W_p(\mu_i{}^\lambda, \mu_i'^\lambda) + W_p(\mu_i'^\lambda, \mu_i'). \tag{C.46}$$

Finally, since we have shown the inequality (C.30) for $1 \leq p < \infty$, the bound for $p = \infty$ follows by Proposition 3 of Givens and Shortt (1984). □

The following result is an elementary fact from linear algebra (proof is included for completeness).

Lemma 1. *Suppose that $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, and $\|\mathbf{v}\| \leq \|\mathbf{u}\|$. Then there exists a symmetric matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ such that $\mathbf{W}\mathbf{u} = \mathbf{v}$, and $-\mathbf{I} \preceq \mathbf{W} \preceq \mathbf{I}$ ($\preceq$ denotes the partial Loewner ordering).*

Proof. First we assume that $\|\mathbf{u}\| = \|\mathbf{v}\|$. If $\mathbf{u} = \mathbf{v}$, then $\mathbf{W} = \mathbf{I}$ works, otherwise it is easy to check that

$$\mathbf{W} = (\mathbf{u}+\mathbf{v})(\mathbf{u}+\mathbf{v})^T/\|\mathbf{u}+\mathbf{v}\|^2 - (\mathbf{u}-\mathbf{v})(\mathbf{u}-\mathbf{v})^T/\|\mathbf{u}-\mathbf{v}\|^2 \tag{C.47}$$

satisfies the requirements. The general case follows by rescaling. □

Now we are ready to prove our contraction bound.

Proof of Theorem 3. Let $(\mathbf{Z}_{1:b}, \mathbf{Z}'_{1:b})$ be a coupling of the two distributions $\pi_\rho(\mathbf{Z}_{1:b}|\theta)$ and $\pi_\rho(\mathbf{Z}'_{1:b}|\theta)$ such that

$$\|\mathbf{Z}_i - \mathbf{Z}_i'\| \leq \frac{1}{1+\rho^2 m_i}\|\mathbf{A}_i(\boldsymbol{\theta}-\boldsymbol{\theta}')\| \text{ almost surely.} \tag{C.48}$$

The existence of such a coupling follows from Proposition 18. Given this coupling $(\mathbf{Z}_{1:b}, \mathbf{Z}'_{1:b})$, our next step is to couple the two conditional distributions

$$\pi_\rho(\theta|\mathbf{Z}_{1:b}) \sim \mathcal{N}(\mu_\theta(\mathbf{Z}_{1:b}), \Sigma_\theta), \tag{C.49}$$

$$\pi_\rho(\theta|\mathbf{Z}'_{1:b}) \sim \mathcal{N}(\mu_\theta(\mathbf{Z}'_{1:b}), \Sigma_\theta), \tag{C.50}$$

where $\Sigma_\theta = \rho^2(\sum_{i=1}^b \mathbf{A}_i^T\mathbf{A}_i)^{-1}$ and $\mu_\theta(\mathbf{z}_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^T\mathbf{A}_i)^{-1}\sum_{i=1}^b \mathbf{A}_i^T\mathbf{z}_i$. Since these two Gaussian distributions have the same covariance matrix, coupling them can be done in a straightforward way, and we can see that for the metric w introduced in the statement of Theorem 3, for every $1 \leq p \leq \infty$, we have

$$W_p^w(\mathbf{P}_{\mathrm{SGS}}(\theta,\cdot), \mathbf{P}_{\mathrm{SGS}}(\theta',\cdot)) \leq \left[\mathbb{E}(w(\mu_\theta(\mathbf{Z}_{1:b}), \mu_\theta(\mathbf{Z}'_{1:b}))^p)\right]^{1/p}, \tag{C.51}$$

where W_p^w denotes Wasserstein distance of order p with respect to the metric w. Note that

$$\mu_\theta(\mathbf{Z}_{1:b}) - \mu_\theta(\mathbf{Z}'_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^T\mathbf{A}_i)^{-1}\sum_{i=1}^b \mathbf{A}_i^T(\mathbf{Z}_i - \mathbf{Z}'_i). \tag{C.52}$$

For each $i \in [b]$, we now apply Lemma 1 with $\mathbf{v} = \mathbf{Z}_i - \mathbf{Z}'_i$ and $\mathbf{u} = \mathbf{A}_i(\boldsymbol{\theta} - \boldsymbol{\theta}')/(1+\rho^2 m_i)$. Using (C.48), the assumption $\|\mathbf{v}\| \leq \|\mathbf{u}\|$ of Lemma 1 is satisfied and there exist some symmetric matrices $\mathbf{W}_1, \ldots, \mathbf{W}_b \in \mathbb{R}^{d\times d}$ such that $-\mathbf{I} \preceq \mathbf{W}_i \preceq \mathbf{I}$, and

$$\mathbf{Z}_i - \mathbf{Z}'_i = \mathbf{W}_i \frac{\mathbf{A}_i(\boldsymbol{\theta} - \boldsymbol{\theta}')}{1+\rho^2 m_i}, \quad \forall i \in [b]. \tag{C.53}$$

This yields

$$\mu_\theta(\mathbf{Z}_{1:b}) - \mu_\theta(\mathbf{Z}'_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^T\mathbf{A}_i)^{-1}\sum_{i=1}^b \frac{\mathbf{A}_i^T\mathbf{W}_i\mathbf{A}_i}{1+\rho^2 m_i}(\boldsymbol{\theta} - \boldsymbol{\theta}') \text{ almost surely.} \tag{C.54}$$

From the definition of w, we can now write

$$w(\mu_\theta(\mathbf{Z}_{1:b}), \mu_\theta(\mathbf{Z}'_{1:b})) = \left\| (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{1/2} (\mu_\theta(\mathbf{Z}_{1:b}) - \mu_\theta(\mathbf{Z}'_{1:b})) \right\| \tag{C.55}$$

$$= \left\| (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{-1/2} \sum_{i=1}^{b} \frac{\mathbf{A}_i^T \mathbf{W}_i \mathbf{A}_i}{1+\rho^2 m_i} (\boldsymbol{\theta} - \boldsymbol{\theta}') \right\| \tag{C.56}$$

$$= \left\| (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{-1/2} \sum_{i=1}^{b} \frac{\mathbf{A}_i^T \mathbf{W}_i \mathbf{A}_i}{1+\rho^2 m_i} (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{-1/2} \cdot (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}') \right\| \tag{C.57}$$

$$\leq \left\| (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{-1/2} \left(\sum_{i=1}^{b} \frac{\mathbf{A}_i^T \mathbf{A}_i}{1+\rho^2 m_i} \right) (\sum_{i=1}^{b} \mathbf{A}_i^T \mathbf{A}_i)^{-1/2} \right\| w(\boldsymbol{\theta}, \boldsymbol{\theta}') \text{ almost surely.} \tag{C.58}$$

Hence the result follows from (C.51). □

C.2 Proof of Corollary 4

First, we will show the convergence results in Wasserstein distance of order p for $1 \leq p < \infty$. Let $(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)$ be the optimal coupling of the initial distribution ν and the stationary distribution π_ρ that achieves the Wasserstein distance of order p for the metric w (see Theorem 4.1 of Villani (2008) for proof of existence), i.e.

$$W_p^w(\nu, \pi_\rho) = \|w(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)\|_{L^p}. \tag{C.59}$$

For $i \geq 1$, assuming that $(\boldsymbol{\theta}_{0:i-1}, \boldsymbol{\theta}'_{0:i-1})$ has been defined, add two more elements $(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i)$ by defining their conditional distribution based on the past elements as the optimal coupling between $\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}_{i-1}, \cdot)$ and $\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}'_{i-1}, \cdot)$ achieving the Wasserstein distance of order p for the metric w. Using that $K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq K_{\mathrm{SGS}}$ by Theorem 3, we have

$$\mathbb{E}(w(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_1)^p | \boldsymbol{\theta}_0, \boldsymbol{\theta}'_0) \leq (1 - K_{\mathrm{SGS}})^p w(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)^p, \tag{C.60}$$

and so by the tower property, we have

$$\|w(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_1)\|_{L^p} \leq (1 - K_{\mathrm{SGS}}) W_p^w(\nu, \pi_\rho). \tag{C.61}$$

Similarly, by induction, it follows that

$$\|w(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i)\|_{L^p} \leq (1 - K_{\mathrm{SGS}})^i W_p^w(\nu, \pi_\rho). \tag{C.62}$$

Now (3.11) for $1 \leq p < \infty$ follows by noticing that $\boldsymbol{\theta}'_i \sim \pi_\rho$ since the Markov chain $(\boldsymbol{\theta}'_j)_{j \geq 0}$ was initialized in its stationary distribution. Finally, the $p = \infty$ case follows from Proposition 3 of Givens and Shortt (1984).

Regarding the convergence rate in total variation distance stated in Theorem 3, we will use Corollary 6 and Proposition 19 detailed below.

Corollary 6 (Lower bound on the spectral gap of SGS)**.** *SGS defines a reversible Markov chain. Under Assumption* (A_3)*, its absolute spectral gap* γ^*_{SGS} *is lower bounded by* K_{SGS}*, see* (3.10)*.*

Proof. The reversibility follows by a standard argument for data augmentation schemes given in Lemma 3.1 of Liu, Wong, and Kong (1994). The lower bound on the absolute spectral gap follows by Proposition 30 of Ollivier (2009). □

The following proposition is well known in the MCMC literature but we have only found a proof for Markov chains on finite state spaces. Hence for completeness, we include a short proof here.

Proposition 19. *Suppose that* $\mathbf{P}(\mathbf{z}, \cdot)$ *is a reversible Markov kernel on a Polish state space* Ω *with absolute spectral gap* $\gamma^* > 0$*, and unique stationary distribution* π*. Then for any initial distribution* ν *that is absolutely continuous with respect to* π*, and any number of steps* $t \in \mathbb{Z}_+$*, we have*

$$\left\|\nu\mathbf{P}^t - \pi\right\|_{\mathrm{TV}} \leq \frac{1}{2}\left(\mathbb{E}_\pi\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)^2\right] - 1\right)^{1/2} \cdot (1-\gamma^*)^t. \tag{C.63}$$

Our proof is based on the following lemma.

Lemma 2. *Suppose that* $\mathbf{Q}(x, \mathrm{d}y)$ *is a reversible Markov kernel on a Polish state space* Ω *with stationary distribution* π*. Then for any distribution* ν *that is absolutely continuous with respect to* π*,* $\nu\mathbf{Q}$ *is also absolutely continuous with respect to* π*, and for* π*-almost every* $x \in \Omega$*, we have*

$$\frac{\mathrm{d}(\nu\mathbf{Q})}{\mathrm{d}\pi}(x) = \left(\mathbf{Q}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x). \tag{C.64}$$

Proof. The claim of the lemma is equivalent to showing that for every bounded measurable function $f : \Omega \to \mathbb{R}$, we have

$$\int_{x\in\Omega} \frac{\mathrm{d}(\nu\mathbf{Q})}{\mathrm{d}\pi}(x) f(x) \pi(\mathrm{d}x) = \int_{x\in\Omega} \left(\mathbf{Q}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x) f(x) \pi(\mathrm{d}x). \tag{C.65}$$

Since if we add a constant to f, both sides increase by this constant, we can assume without loss of generality that f is non-negative. Under this

assumption, we have

$$\int_{x\in\Omega}\frac{\mathrm{d}(\nu\mathbf{Q})}{\mathrm{d}\pi}(x)f(x)\pi(\mathrm{d}x)=\int_{x\in\Omega}f(x)(\nu\mathbf{Q})(\mathrm{d}x) \quad (C.66)$$

$$=\int_{x,y\in\Omega}f(x)\nu(\mathrm{d}y)\mathbf{Q}(y,\mathrm{d}x) \quad (C.67)$$

$$=\int_{x,y\in\Omega}f(y)\nu(\mathrm{d}x)\mathbf{Q}(x,\mathrm{d}y) \quad (C.68)$$

$$=\int_{x,y\in\Omega}f(y)\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x)\pi(\mathrm{d}x)\mathbf{Q}(x,\mathrm{d}y) \quad (C.69)$$

by the monotone convergence theorem (using the non-negativity of f)

$$=\lim_{M\to\infty}\int_{x,y\in\Omega}f(y)\min\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x),M\right)\pi(\mathrm{d}x)\mathbf{Q}(x,\mathrm{d}y) \quad (C.70)$$

using the reversibility of $\mathbf{Q}$ (in the equivalent bounded measurable test function formulation)

$$=\lim_{M\to\infty}\int_{x,y\in\Omega}f(y)\min\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x),M\right)\pi(\mathrm{d}y)\mathbf{Q}(y,\mathrm{d}x) \quad (C.71)$$

by the monotone convergence theorem (using the non-negativity of f)

$$=\int_{x,y\in\Omega}f(y)\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x)\pi(\mathrm{d}y)\mathbf{Q}(y,\mathrm{d}x) \quad (C.72)$$

$$=\int_{y\in\Omega}f(y)\left(\mathbf{Q}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(y)\pi(\mathrm{d}y), \quad (C.73)$$

hence (C.65) and the claim of our lemma holds. □

Proof of Proposition 19. We define the Hilbert space $L^2(\pi)$ as measurable functions f on Ω satisfying $\mathbb{E}_\pi(f^2)<\infty$, endowed with the scalar product $\langle f,g\rangle_\pi=\int_{\mathbf{z}\in\Omega}f(\mathbf{z})g(\mathbf{z})\pi(\mathrm{d}\mathbf{z})$. Let us define the linear operator $\mathbf{\Pi}(f)(\mathbf{z}):=\mathbb{E}_\pi(f)$ for any $f\in L^2(\pi)$, $\mathbf{z}\in\Omega$.

Using Lemma 2 with $\mathbf{Q}=\mathbf{P}^t$, it follows that

$$\|\nu\mathbf{P}^t-\pi\|_{\mathrm{TV}}=\frac{1}{2}\int_{x\in\Omega}\left|\frac{\mathrm{d}\nu\mathbf{P}^t}{\mathrm{d}\pi}(x)-1\right|\pi(\mathrm{d}x)$$

using Jensen's inequality, we have

$$\leq\frac{1}{2}\sqrt{\int_{x\in\Omega}\left(\frac{\mathrm{d}\nu\mathbf{P}^t}{\mathrm{d}\pi}(x)-1\right)^2\pi(\mathrm{d}x)}. \quad (C.74)$$

Using Lemma 2 again, the integral inside the square root can be further

bounded as

$$\begin{aligned}
\int_{x\in\Omega}\left(\frac{\mathrm{d}\nu\mathbf{P}^t}{\mathrm{d}\pi}(x)-1\right)^2\pi(\mathrm{d}x) &= \int_{x\in\Omega}\left(\left(\mathbf{P}^t\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x)-1\right)^2\pi(\mathrm{d}x)\\
&= \int_{x\in\Omega}\left(\left((\mathbf{P}^t-\mathbf{\Pi})\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x)\right)^2\pi(\mathrm{d}x)\\
&= \int_{x\in\Omega}\left(\left((\mathbf{P}-\mathbf{\Pi})^t\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x)\right)^2\pi(\mathrm{d}x)\\
&= \int_{x\in\Omega}\left(\left((\mathbf{P}-\mathbf{\Pi})^t\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right)\right)(x)\right)^2\pi(\mathrm{d}x)\\
&= \left\langle\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1,(\mathbf{P}-\mathbf{\Pi})^{2t}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right)\right\rangle_\pi\\
&\le \|\mathbf{P}-\mathbf{\Pi}\|_\pi^{2t}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right\|_\pi^2\\
&= (1-\gamma^*)^{2t}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right\|_\pi^2,
\end{aligned}$$

and the claim of the proposition follows by substituting this into (C.74).

□

Now we are ready to prove our convergence bound in total variation distance. From Corollary 6, we know that the absolute spectral gap of SGS satisfies that $\gamma^* \ge K_{\mathrm{SGS}}$ (defined in (3.10)), and Proposition 19 implies that

$$\left\|\nu\mathbf{P}_{\mathrm{SGS}}^t-\pi_\rho\right\|_{\mathrm{TV}} \le \sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right]-1}\cdot(1-\gamma^*)^t \tag{C.75}$$

$$\le \sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right]-1}\cdot(1-K_{\mathrm{SGS}})^t. \tag{C.76}$$

C.3 Proof of Theorem 4

Proof of Theorem 4. From Proposition 7, it follows that if ρ is chosen as in (3.16), then

$$W_1(\pi_\rho,\pi) \le \frac{\epsilon}{2}\cdot\frac{\sqrt{d}}{\sqrt{m_1}}. \tag{C.77}$$

From Proposition 1 part (ii) in the work by Durmus and Moulines (2019) it follows that for the initial distribution $\delta_{\boldsymbol{\theta}^*}$ (Dirac measure at $\boldsymbol{\theta}^*$), we have

$$W_1(\delta_{\boldsymbol{\theta}^*},\pi) \le W_2(\delta_{\boldsymbol{\theta}^*},\pi) \le \frac{\sqrt{d}}{\sqrt{m_1}}, \tag{C.78}$$

and hence by combining this with (C.77) using the triangle inequality and the assumption $\epsilon \leq 1$, it follows that

$$W_1(\delta_{\boldsymbol{\theta}^*}, \pi_\rho) \leq \frac{3}{2}\frac{\sqrt{d}}{\sqrt{m_1}}. \tag{C.79}$$

Now from Theorem 3, it follows that the coarse Ricci curvature of SGS is lower bounded by

$$K_{\mathrm{SGS}} := \frac{\rho^2 m_1}{1+\rho^2 m_1}, \tag{C.80}$$

and therefore by Corollary 21 of Ollivier (2009), we have

$$W_1(P^t_{\mathrm{SGS}}(\boldsymbol{\theta}^*, \cdot), \pi_\rho) \leq W_1(\delta_{\boldsymbol{\theta}^*}, \pi_\rho) \cdot (1 - K_{\mathrm{SGS}})^t \leq \frac{\epsilon}{2} \cdot \frac{\sqrt{d}}{\sqrt{m_1}}.$$

The claim of the theorem now follows by the triangle inequality. □

C.4 Bounds for SGS with rejection sampling

The following bound is a standard result in rejection sampling (see for instance Section 2.3 of Robert and Casella (2004)).

Lemma 3. *Suppose that $\mu(\mathbf{z}) = \tilde{\mu}(\mathbf{z})/\tilde{Z}$ is the target density on $\mathbb{R}^d$, and $\nu(\mathbf{z})$ is the proposal density (both absolutely continuous w.r.t. the Lebesgue measure). Here $\tilde{\mu}(\mathbf{z})$ is the unnormalized target and $\tilde{Z}$ is the normalising constant (which is typically unknown). Suppose that the condition*

$$\tilde{\mu}(\mathbf{z}) \leq M\nu(\mathbf{z})$$

holds for some constant $M < \infty$ for every $\mathbf{z} \in \mathbb{R}^d$. Under this assumption, if we take samples $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ from ν and accept $\mathbf{Z}_i$ with probability $\frac{\tilde{\mu}(\mathbf{Z}_i)}{M\nu(\mathbf{Z}_i)}$, then the accepted samples will be distributed according to μ. Moreover, the expected number of samples taken until the first acceptance is equal to $M/\tilde{Z}$.

The following lemma gives a complexity bound for rejection sampling for log-concave distributions. We assume that we have access to an approximation of the minimum of the strongly convex and smooth potential U, which will be denoted by $\tilde{\mathbf{z}}$. The quality of this approximation is taken into account in the proposal distribution using the norm of $\nabla U(\tilde{\mathbf{z}})$.

Lemma 4. *Suppose that $\mu(\mathbf{z}) \propto \exp(-U(\mathbf{z}))$ is a distribution on $\mathbb{R}^d$ such that U is twice differentiable and*

$$A\mathbf{I}_d \preceq \nabla^2 U(\mathbf{z}) \preceq B\mathbf{I}_d \tag{C.81}$$

for some $0 < A \leq B$ (strongly convex and smooth). Let $\mathbf{z}^*$ be the

unique minimizer of U, $\tilde{\mathbf{z}}$ *another point (an approximation of* $\mathbf{z}^*$*), and* $\nu(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \tilde{\mathbf{z}}, \tilde{A}^{-1}\mathbf{I}_d)$, *where*

$$\tilde{A} = A + \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2d} - \sqrt{\frac{\|\nabla U(\tilde{\mathbf{z}})\|^4}{4d^2} + \frac{A\|\nabla U(\tilde{\mathbf{z}})\|^2}{d}}. \tag{C.82}$$

Suppose that we take samples $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ *from* ν, *and accept them with probability*

$$\mathbb{P}(\mathbf{Z}_j \text{ is accepted}) = \exp\left(-\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})} - [U(\mathbf{z}) - U(\tilde{\mathbf{z}}))] + \frac{\tilde{A}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2}{2}\right). \tag{C.83}$$

Then these accepted samples are distributed according to μ. *Moreover, the expected number of samples taken until one is accepted is less than or equal to*

$$\left(B/\tilde{A}\right)^{d/2} \cdot \exp\left[\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2}\left(\frac{1}{A-\tilde{A}} - \frac{1}{B}\right)\right]. \tag{C.84}$$

Proof. The proposal density equals

$$\nu(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \tilde{\mathbf{z}}, \tilde{A}^{-1}\mathbf{I}_d) \tag{C.85}$$

$$= \exp\left(-\frac{\tilde{A}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2}{2}\right) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2}. \tag{C.86}$$

We define the unnormalized version of μ as

$$\tilde{\mu}(\mathbf{z}) = \exp(-[U(\mathbf{z}) - U(\tilde{\mathbf{z}})]) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2}. \tag{C.87}$$

Notice that

$$U(\mathbf{z}) - U(\tilde{\mathbf{z}}) = \left\langle \int_{t=0}^{1} \nabla U(\tilde{\mathbf{z}} + t(\mathbf{z}-\tilde{\mathbf{z}}))\mathrm{d}t, \mathbf{z}-\tilde{\mathbf{z}} \right\rangle. \tag{C.88}$$

By the intermediate value theorem, there is some $\mathbf{z}(t)$ such that

$$= \left\langle \nabla U(\tilde{\mathbf{z}}), \mathbf{z}-\tilde{\mathbf{z}} \right\rangle + \left\langle \mathbf{z}-\tilde{\mathbf{z}}, \left(\int_{t=0}^{1} t\nabla^2 U(\mathbf{z}(t))\mathrm{d}t\right)^T (\mathbf{z}-\tilde{\mathbf{z}}) \right\rangle, \tag{C.89}$$

so using the assumption (C.81) it follows that

$$\geq -\|\nabla U(\tilde{\mathbf{z}})\|\|\mathbf{z}-\tilde{\mathbf{z}}\| + \frac{A}{2}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2. \tag{C.90}$$

Based on this, one gets

$$\frac{\tilde{\mu}(\mathbf{z})}{\nu(\mathbf{z})} \leq \exp\left(\|\nabla U(\tilde{\mathbf{z}})\| \cdot \|\mathbf{z}-\tilde{\mathbf{z}}\| - \frac{A-\tilde{A}}{2}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2\right) \leq \exp\left(\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})}\right). \tag{C.91}$$

Hence we have $\tilde{\mu}(\mathbf{z}) \leq M\nu(\mathbf{z})$ for $M = \exp\left(\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})}\right)$.

For the normalising constant, we have

$$\tilde{Z} = \int_{\mathbf{z}\in\mathbb{R}^d} \tilde{\mu}(\mathbf{z})\mathrm{d}\mathbf{z} \tag{C.92}$$

$$= \exp(U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*)) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2} \cdot \int_{\mathbf{z}\in\mathbb{R}^d} \exp(-(U(\mathbf{z}) - U(\mathbf{z}^*)))\mathrm{d}\mathbf{z} \tag{C.93}$$

using Taylor's expansion with second order remainder term, and assumption (C.81) yields

$$\geq \exp(U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*)) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2} \cdot \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(-\frac{B}{2}\|\mathbf{z}-\mathbf{z}^*\|^2\right)\mathrm{d}\mathbf{z} \tag{C.94}$$

$$= \left(\frac{\tilde{A}}{B}\right)^{d/2} \cdot \exp\left(U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*)\right) \tag{C.95}$$

$$\geq \left(\frac{\tilde{A}}{B}\right)^{d/2} \exp\left(\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2B}\right), \tag{C.96}$$

where in the last step we have used the fact that for $\mathbf{z}' = \tilde{\mathbf{z}} - \frac{\|\nabla U(\tilde{\mathbf{z}})\|}{B}$, we have

$$U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*) \geq U(\tilde{\mathbf{z}}) - U(\mathbf{z}') \tag{C.97}$$

$$= \left\langle \int_{t=0}^{1} \nabla U(\tilde{\mathbf{z}} + t(\mathbf{z}'-\tilde{\mathbf{z}}))\mathrm{d}t, \tilde{\mathbf{z}} - \mathbf{z}' \right\rangle \tag{C.98}$$

using the fact that $\mathbf{z}^*$ is the minimum of U.

By the intermediate value theorem, there is some $\tilde{\mathbf{z}}(t) \in \mathbb{R}^d$ such that

$$= \left\langle \nabla U(\tilde{\mathbf{z}}), \tilde{\mathbf{z}} - \mathbf{z}' \right\rangle + \left\langle \tilde{\mathbf{z}} - \mathbf{z}', \left(\int_{t=0}^{1} t\nabla^2 U(\tilde{\mathbf{z}}(t))\mathrm{d}t\right) \cdot (\tilde{\mathbf{z}} - \mathbf{z}') \right\rangle \tag{C.99}$$

$$\geq \left\langle \nabla U(\tilde{\mathbf{z}}), \tilde{\mathbf{z}} - \mathbf{z}' \right\rangle - \frac{B}{2}\|\tilde{\mathbf{z}} - \mathbf{z}'\|^2 = \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2B}. \tag{C.100}$$

Now it follows by Lemma 3 and the above bound on $\tilde{Z}$ that the expected number of samples until the first acceptance is less than or equal to

$$E(\tilde{A}) := \exp\left(\|\nabla U(\tilde{\mathbf{z}})\|^2 \left(\frac{1}{2(A-\tilde{A})} - \frac{1}{2B}\right)\right)\left(\frac{B}{\tilde{A}}\right)^{d/2}. \quad \text{(C.101)}$$

The parameter $\tilde{A}$ in (C.82) is chosen such that $E(\tilde{A})$ is minimized. Note that the minimizer of $E(\tilde{A})$ is the same as the minimizer of

$$\log(E(\tilde{A})) = \frac{d}{2}\log(B) - \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2B} + \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})} - \frac{d}{2}\log(\tilde{A}). \quad \text{(C.102)}$$

It is easy to check that this is a strictly convex function of $\tilde{A}$ on the interval $(0, A)$, and hence the unique minimum is taken at a point where the derivative is zero. This point, denoted by $\tilde{A}_{\min}$, thus satisfies

$$\left.\frac{\partial \log(E(\tilde{A}))}{\partial \tilde{A}}\right|_{\tilde{A}=\tilde{A}_{\min}} = \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})^2} - \frac{d}{2}\cdot\frac{1}{\tilde{A}} = 0. \quad \text{(C.103)}$$

Hence by rearrangement

$$(\tilde{A}-A)^2 - (\|\nabla U(\tilde{\mathbf{z}})\|^2/d)\tilde{A} = 0 \quad \text{(C.104)}$$

$$\tilde{A}^2 - (2A + \|\nabla U(\tilde{\mathbf{z}})\|^2/d)\tilde{A} + A^2 = 0 \quad \text{(C.105)}$$

$$\tilde{A} = \frac{(2A + \|\nabla U(\tilde{\mathbf{z}})\|^2/d) \pm \sqrt{(2A + \|\nabla U(\tilde{\mathbf{z}})\|^2/d)^2 - 4A^2}}{2} \quad \text{(C.106)}$$

$$= A + \|\nabla U(\tilde{\mathbf{z}})\|^2/(2d) \pm \sqrt{\|\nabla U(\tilde{\mathbf{z}})\|^4/(4d^2) + A\|\nabla U(\tilde{\mathbf{z}})\|^2/d}. \quad \text{(C.107)}$$

Only the solution with the $-$ sign falls in the interval $(0, A)$, hence it is the minimizer of $M/\tilde{Z}$. □

Corollary 7 (Complexity of rejection sampling for sampling $\mathbf{z}_i$ given $\boldsymbol{\theta}$)**.** *Suppose that Assumptions (A_1) and (A_2) (smoothness and convexity) hold, and that f_i is m_i-strongly convex for some $m_i \geq 0$ (possibly zero). Let*

$$U_i(\mathbf{z}_i) := f_i(\mathbf{z}_i) + \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2}{2\rho^2},$$

$\mathbf{z}_i^*(\boldsymbol{\theta})$ be the unique minimizer of U_i, and $\tilde{\mathbf{z}}_i(\boldsymbol{\theta})$ be another point (an approximation of $\mathbf{z}_i^*(\boldsymbol{\theta})$). We let

$$\tilde{A}_i = \frac{1}{\rho^2} + m_i + \frac{\|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{2d_i} - \sqrt{\frac{\|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^4}{4d_i^2} + \frac{(1/\rho^2 + m_i)\,\|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{d_i}},$$

and set $\nu_{\boldsymbol{\theta}}(\mathbf{z}_i) := \mathcal{N}(\mathbf{z}_i; \tilde{\mathbf{z}}_i(\boldsymbol{\theta}), (\tilde{A}_i)^{-1} \cdot \mathbf{I}_{d_i})$.

Suppose that we take samples $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ *from* $\nu_{\boldsymbol{\theta}}$*, and accept them with probability*

$$\mathbb{P}(\mathbf{Z}_j \text{ is accepted}) = \exp\left(-\frac{\|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{2(1/\rho^2 + m_i - \tilde{A}_i)} - [U_i(\mathbf{Z}_j) - U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))] + \frac{\tilde{A}_i \|\mathbf{Z}_j - \tilde{\mathbf{z}}_i(\boldsymbol{\theta})\|^2}{2} \right).$$

Then these accepted samples are distributed according to $\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$*. Moreover, the expected number of samples taken until one is accepted is equal to*

$$E_i := \left(\frac{1/\rho^2 + M_i}{\tilde{A}_i} \right)^{d_i/2} \cdot \exp\left[\frac{\|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{2} \left(\frac{1}{1/\rho^2 + m_i - \tilde{A}_i} - \frac{1}{1/\rho^2 + M_i} \right) \right], \tag{C.108}$$

which is less than or equal to 2 if

$$\rho^2(2d_i(M_i - m_i) - m_i) \le 1 \quad \text{and} \quad \|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\| \le \frac{2}{7} \cdot \frac{\sqrt{1/\rho^2 + m_i}}{\sqrt{d_i}}. \tag{C.109}$$

Remark 2. *The choice of the approximate minimizer* $\tilde{\mathbf{z}}_i(\boldsymbol{\theta})$ *that we are using in our implementation is a few steps of gradient descent started from* $\tilde{\mathbf{z}}_i^{(0)}(\boldsymbol{\theta}) = \mathbf{A}_i\boldsymbol{\theta}$*, with step size* $\frac{1}{1/\rho^2 + M_i}$*, i.e. for* $j \ge 1$*,*

$$\tilde{\mathbf{z}}_i^{(j)}(\boldsymbol{\theta}) = \tilde{\mathbf{z}}_i^{(j-1)}(\boldsymbol{\theta}) - \nabla U_i(\tilde{\mathbf{z}}_i^{(j-1)}(\boldsymbol{\theta})) \cdot \frac{1}{1/\rho^2 + M_i}.$$

We stop once the condition $\|\nabla U_i(\tilde{\mathbf{z}}_i^{(j)}(\boldsymbol{\theta}))\| \le \frac{2}{7} \cdot \frac{\sqrt{1/\rho^2 + m_i}}{\sqrt{d_i}}$ *is satisfied, and set* $\tilde{\mathbf{z}}_i$ *to* $\tilde{\mathbf{z}}_i^{(j)}$*.*

Since the condition number of the function U_i *equals* $\kappa_i = \frac{1+\rho^2 M_i}{1+\rho^2 m_i}$*, and the gradient descent decreases the norm of the gradient by a factor of* $1 - 1/\kappa_i$ *in each iteration, it follows that we need at most*

$$\left\lceil \frac{\log \|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\| - \log\left(\frac{2}{7} \cdot \frac{\sqrt{1/\rho^2 + m_i}}{\sqrt{d_i}} \right)}{\log(1/(1 - 1/\kappa_i))} \right\rceil$$

iterations before stopping.

Proof of Corollary 7. The fact that the accepted samples are distributed according to $\pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ and the formula (C.108) about the expected number of samples until acceptance follows from Lemma 4.

Let $G := \|\nabla U_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|$, then $\tilde{A}_i = 1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}$,

and we have

$$\log(E_i) = \frac{d_i}{2}\log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) + \frac{G^2}{2}\left(\frac{1}{\sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i} - G^2/(2d_i)} - \frac{1}{1/\rho^2 + M_i}\right). \tag{C.110}$$

For the first part, notice that

$$\begin{aligned}
&\log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) \\
&= \log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i}\right) + \log\left(\frac{1/\rho^2 + m_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) \\
&= \log\left(1 + \frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i}\right) + \log\left(\frac{1}{1 + c - \sqrt{c^2 + 2c}}\right),
\end{aligned}$$

where $c = \frac{G^2/(2d_i)}{1/\rho^2 + m_i}$. Now using the fact that $\log(1+x) \leq x$ for $x > 0$, and that $\log\left(\frac{1}{1+c-\sqrt{c^2+2c}}\right) \leq \sqrt{2c}$ for $c \geq 0$, it follows that we have

$$\begin{aligned}
&\frac{d_i}{2}\log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) \\
&\leq \frac{d_i}{2}\left(\frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i} + \frac{G}{\sqrt{d_i(1/\rho^2 + m_i)}}\right).
\end{aligned}$$

For the second part (C.110),

$$\begin{aligned}
&\frac{G^2}{2}\left(\frac{1}{\sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i} - G^2/(2d_i)} - \frac{1}{1/\rho^2 + M_i}\right) \\
&= \frac{d_i}{\sqrt{1 + 4\left(1/\rho^2 + m_i\right)d_i/G^2} - 1} - \frac{G^2}{2}\cdot\frac{1}{1/\rho^2 + M_i}
\end{aligned}$$

using the fact that $\frac{1}{\sqrt{1+x}-1} \leq \frac{2}{\sqrt{x}}$ for $x \geq 2$, for $G \leq \sqrt{2d_i(1/\rho^2 + m_i)}$, we have

$$\leq G\cdot\frac{\sqrt{d_i}}{\sqrt{1/\rho^2 + m_i}} - \frac{G^2}{2}\cdot\frac{1}{1/\rho^2 + M_i}.$$

Hence by combining these terms, we obtain that for $G \leq \sqrt{2d_i(1/\rho^2 + m_i)}$,

$$\log(E_i) \leq \frac{d_i}{2}\frac{\rho^2(M_i - m_i)}{1+\rho^2 m_i} + G \cdot \frac{3}{2} \cdot \frac{\sqrt{d_i}}{\sqrt{(1/\rho^2 + m_i)}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2 + M_i}$$

Under the first part of assumption (C.109), $\rho^2(2d_i(M_i - m_i) - m_i) \leq 1$, one can check that $\frac{d_i}{2}\frac{\rho^2(M_i-m_i)}{1+\rho^2 m_i} \leq \frac{1}{4}$. Using the second part of (C.109), $G \leq \frac{2}{7} \cdot \frac{\sqrt{1/\rho^2+m_i}}{\sqrt{d_i}}$, it follows that $G \cdot \frac{3}{2} \cdot \frac{\sqrt{d_i}}{\sqrt{(1/\rho^2+m_i)}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2+M_i} \leq \log(2) - \frac{1}{4}$, so $\log(E_i) \leq \log(2)$ and our claim holds. □

C.5 Proof of Theorem 5

The next two lemmas will be used for obtaining our total variation distance convergence rates.

Lemma 5. *Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and M-gradient-Lipschitz. Then for every $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\|\nabla U(\mathbf{x})\|^2 \leq 2M(U(\mathbf{x}) - \inf_{\mathbf{x}\in\mathbb{R}^d} U(\mathbf{x})).$$

Proof. Let $\mathbf{x}' = \mathbf{x} - \nabla U(\mathbf{x})/M$, then we have

$$\begin{aligned} U(\mathbf{x}) - U(\mathbf{x}') &= \int_{t=0}^{1} \langle \nabla U(\mathbf{x} + t(\mathbf{x}' - \mathbf{x})), \mathbf{x} - \mathbf{x}' \rangle \, \mathrm{d}t \\ &= \langle \nabla U(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle + \int_{t=0}^{1} \langle \nabla U(\mathbf{x} + t(\mathbf{x}' - \mathbf{x})) - \nabla U(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle \, \mathrm{d}t \end{aligned}$$

using the M-gradient Lipschitz property

$$\begin{aligned} &\geq \langle \nabla U(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle + \int_{t=0}^{1} Mt\|\mathbf{x} - \mathbf{x}'\|^2 \mathrm{d}t \\ &\geq \frac{\|\nabla U(\mathbf{x})\|^2}{2M}, \end{aligned}$$

hence the result. □

Lemma 6. *Suppose that Assumptions (A_0), (A_1) and (A_3) hold. Assume also that $b = 1$, $d_1 = d$ and $\mathbf{A}_1$ has full rank. Let $\boldsymbol{\theta}^*$ be the minimizer of $f(\boldsymbol{\theta}) = f_1(\mathbf{A}_1\boldsymbol{\theta})$, and $\nu(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, (M_1\mathbf{A}_1^T\mathbf{A}_1)^{-1})$. Then, for any $\rho > 0$, we have $\frac{\nu(\boldsymbol{\theta})}{\pi_\rho(\boldsymbol{\theta})} \leq C_\rho$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$, where*

$$C_\rho := (1 + \rho^2 M_1)^{d/2} \left(\frac{M_1}{m_1}\right)^{d/2}. \tag{C.111}$$

Proof. Let

$$f_1^\rho(\mathbf{A}_1\boldsymbol{\theta}) := -\log \int_{\mathbf{z}_1 \in \mathbb{R}^d} \exp\left(-f_1(\mathbf{z}_1) - \frac{\|\mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta}\|^2}{2\rho^2}\right) \cdot \frac{\mathrm{d}\mathbf{z}_1}{(2\pi\rho^2)^{d/2}}.$$

The, we can see that the negative log-likelihood of $\pi_\rho(\boldsymbol{\theta})$ can be written as

$$f^\rho(\boldsymbol{\theta}) := f_1^\rho(\mathbf{A}_1\boldsymbol{\theta})$$

and the corresponding target distribution is thus

$$\pi_\rho(\boldsymbol{\theta}) = \frac{\exp(-f^\rho(\boldsymbol{\theta}))}{Z_{\pi_\rho}},$$

for a normalising constant Z_{π_ρ}. Using the assumptions (A_1) and (A_3) (smoothness and strong convexity), we have

$$f_1(\mathbf{z}_1) \geq f_1(\mathbf{A}_1\boldsymbol{\theta}) + \left\langle \nabla f_1(\mathbf{A}_1\boldsymbol{\theta}), \mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta} \right\rangle + \frac{m_1}{2}\left\|\mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta}\right\|^2 \tag{C.112}$$

$$f_1(\mathbf{z}_1) \leq f_1(\mathbf{A}_1\boldsymbol{\theta}) + \left\langle \nabla f_1(\mathbf{A}_1\boldsymbol{\theta}), \mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta} \right\rangle + \frac{M_1}{2}\left\|\mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta}\right\|^2. \tag{C.113}$$

Using (C.113), we have

$$\begin{aligned}
\exp(-f_1^\rho(\mathbf{A}_1\boldsymbol{\theta})) &= \int_{\mathbf{z}_1\in\mathbb{R}^d} \exp\left(-f_1(\mathbf{z}_1) - \frac{\|\mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta}\|^2}{2\rho^2}\right)\cdot\frac{\mathrm{d}\mathbf{z}_1}{(2\pi\rho^2)^{d/2}} \\
&\geq \frac{\exp\left(-f_1(\mathbf{A}_1\boldsymbol{\theta})\right)}{(2\pi\rho^2)^{d/2}}\int_{\mathbf{z}_1\in\mathbb{R}^d}\exp\left(-\left\langle\nabla f_1(\mathbf{A}_1\boldsymbol{\theta}), \mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta}\right\rangle - \left(\frac{1+M_1\rho^2}{2\rho^2}\right)\left\|\mathbf{z}_1 - \mathbf{A}_1\boldsymbol{\theta}\right\|^2\right)\mathrm{d}\mathbf{z}_1 \\
&= \frac{\exp\left(-f_1(\mathbf{A}_1\boldsymbol{\theta})\right)}{(2\pi\rho^2)^{d/2}}\int_{\mathbf{u}_1\in\mathbb{R}^d}\exp\left(-\left\langle\nabla f_1(\mathbf{A}_1\boldsymbol{\theta}), \mathbf{u}_1\right\rangle - \left(\frac{1+M_1\rho^2}{2\rho^2}\right)\left\|\mathbf{u}_1\right\|^2\right)\mathrm{d}\mathbf{u}_1 \\
&= \frac{\exp\left(-f_1(\mathbf{A}_1\boldsymbol{\theta})\right)}{(2\pi\rho^2)^{d/2}}\int_{\mathbf{u}_1\in\mathbb{R}^d}\exp\left(-\left(\frac{1+M_1\rho^2}{2\rho^2}\right)\left\|\mathbf{u}_1 + \frac{\rho^2\nabla f_1(\mathbf{A}_1\boldsymbol{\theta})}{1+M_1\rho^2}\right\|^2 + \frac{\rho^2\left\|\nabla f_1(\mathbf{A}_1\boldsymbol{\theta})\right\|^2}{2(1+\rho^2M_1)}\right)\mathrm{d}\mathbf{u}_1 \\
&\geq \frac{\exp\left(-f_1(\mathbf{A}_1\boldsymbol{\theta})\right)}{(1+\rho^2M_1)^{d/2}}\cdot\exp\left(\frac{\rho^2\left\|\nabla f_1(\mathbf{A}_1\boldsymbol{\theta})\right\|^2}{2(1+\rho^2M_1)}\right).
\end{aligned} \tag{C.114}$$

Let $f(\boldsymbol{\theta}) = f_1(\mathbf{A}_1\boldsymbol{\theta})$, then from (C.114), it follows that

$$\begin{aligned}
\pi_\rho(\boldsymbol{\theta}) &= \frac{\exp(-f^\rho(\boldsymbol{\theta}))}{Z_{\pi_\rho}} \\
&\geq \frac{\exp(-f(\boldsymbol{\theta}))}{Z_{\pi_\rho}}\cdot\frac{1}{(1+\rho^2M_1)^{d/2}} \\
&\geq \frac{\exp\left(-f(\boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T(M_1\mathbf{A}_1^T\mathbf{A}_1)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right)}{Z_{\pi_\rho}}\cdot\frac{1}{(1+\rho^2M_1)^{d/2}}.
\end{aligned} \tag{C.115}$$

To lower bound $\pi_\rho(\boldsymbol{\theta})$, we need to upper bound Z_{π_ρ}. Since $Z_{\pi_\rho} = Z_\pi$,

we can do this based on an upper bound on Z_π. Using (A_3), we have

$$\begin{aligned} Z_\pi &= \int_{\mathbb{R}^d} \exp\left(-f_1(\mathbf{A}_1\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} \\ &\leq \exp\left(-f_1(\mathbf{A}_1\boldsymbol{\theta}^*)\right) \int_{\mathbb{R}^d} \exp\left(-\frac{m_1}{2}\left\|\mathbf{A}_1\boldsymbol{\theta} - \mathbf{A}_1\boldsymbol{\theta}^*\right\|^2\right) \mathrm{d}\boldsymbol{\theta} \\ &= \exp\left(-f(\boldsymbol{\theta}^*)\right) (2\pi)^{d/2} \det\left(m_1 \mathbf{A}_1^T \mathbf{A}_1\right)^{-1/2}. \end{aligned} \tag{C.116}$$

By substituting this into (C.115), we obtain that

$$\pi_\rho(\boldsymbol{\theta}) \geq \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T (M_1 \mathbf{A}_1^T \mathbf{A}_1)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right) \cdot \frac{\det\left(m_1 \mathbf{A}_1^T \mathbf{A}_1\right)^{1/2}}{(2\pi)^{d/2}(1+\rho^2 M_1)^{d/2}}.$$

Now the claim of the lemma follows by comparing this with

$$\begin{aligned} \nu(\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^*, (M_1 \mathbf{A}_1^T \mathbf{A}_1)^{-1}) \\ &= \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T (M_1 \mathbf{A}_1^T \mathbf{A}_1)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right) \frac{\det\left(M_1 \mathbf{A}_1^T \mathbf{A}_1\right)^{1/2}}{(2\pi)^{d/2}}. \end{aligned}$$

□

Now we are ready to prove our convergence bound in total variation distance.

Proof of Theorem 5. From Theorem 2, we have

$$\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq \frac{d}{2} M_1 \rho^2.$$

Then, a sufficient condition to satisfy $\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq \epsilon/2$ is to have

$$\rho^2 \leq \frac{\epsilon}{d M_1}. \tag{C.117}$$

From Corollary 6, we know that the absolute spectral gap of SGS satisfies that $\gamma^* \geq K_{\mathrm{SGS}}$ (defined in (3.10)), and Proposition 19 implies that

$$\begin{aligned} \left\|\nu \mathbf{P}_{\mathrm{SGS}}^t - \pi_\rho\right\|_{\mathrm{TV}} &\leq \sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right] - 1} \cdot (1-\gamma^*)^t \\ &\leq \sqrt{\mathbb{E}_{\nu}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)} \cdot (1-K_{\mathrm{SGS}})^t \\ &\leq \sqrt{C_\rho}(1-K_{\mathrm{SGS}})^t, \end{aligned}$$

where in the last step we have used Lemma 6 (C_ρ is defined as in (C.111)). By some algebra, using the definition of $t_{\mathrm{mix}}(\epsilon;\nu)$, and the fact that

$\frac{1}{\log(1/(1-x))} \leq \frac{1}{x}$ for $0 < x < 1$, the above bound implies that

$$\left\| \nu \mathbf{P}_{\mathrm{SGS}}^{t(\epsilon)} - \pi_\rho \right\|_{\mathrm{TV}} \leq \frac{\epsilon}{2},$$

with the choice

$$t \geq \frac{\log\left(\frac{2}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}}. \tag{C.118}$$

Here

$$C = \frac{5d}{8} + \frac{d}{2} \log\left(\frac{M_1}{m_1}\right).$$

With the above choice for ρ^2 and the condition (C.118), the claim of Theorem 5 then follows by the triangle inequality. □

C.6 Details for the toy Gaussian example

This section gives additional details concerning the results depicted on Figure 3.1. For each splitting strategy associated to the model (3.13), we give explicit formulas for the bounds on both TV and 1-Wasserstein distances.

Splitting strategy 1 – Starting from an initial value $\theta_0 \sim \nu$, we now show the explicit form of the Markov transition kernel νP_{SGS}^t after t iterations. To this purpose, we take advantage that the θ-chain corresponds in this case to an auto-regressive process of order 1. Indeed, the conditional distributions of θ and $z_{1:b}$ writing

$$\pi_\rho(z_i|\theta) = \mathcal{N}\left(z_i; \frac{\mu\rho^2 + \theta\sigma^2}{\sigma^2 + \rho^2}, \frac{\rho^2\sigma^2}{\rho^2 + \sigma^2}\right), \forall i \in [b]$$

$$\pi_\rho(\theta|z_{1:b}) = \mathcal{N}\left(\theta; \bar{z}, \frac{\rho^2}{b}\right), \text{ where } \bar{z} := \frac{1}{b}\sum_{i=1}^{b} z_i,$$

we have

$$P_{\mathrm{SGS}} := \Pr\left(\theta^{[t]}|\theta^{[t-1]}\right) = \mathcal{N}\left(\theta^{[t]}; \frac{\sigma^2}{\sigma^2 + \rho^2}\theta^{[t-1]} + \frac{\rho^2}{\sigma^2 + \rho^2}\mu, \frac{2\rho^2\sigma^2 + \rho^4}{b(\rho^2 + \sigma^2)}\right).$$

By a straightforward induction, it follows that the Markov transition kernel νP^t after t iterations and with initial distribution ν has the form

$$\nu P_{\mathrm{SGS}}^t := \Pr\left(\theta^{[t]}|\theta^{[0]} \sim \nu\right)$$
$$= \mathcal{N}\left(\theta^{[t]}; \left(\frac{\sigma^2}{\sigma^2 + \rho^2}\right)^t \theta^{[0]} + \frac{\rho^2\mu}{\sigma^2 + \rho^2}\sum_{i=0}^{t-1}\left(\frac{\sigma^2}{\sigma^2 + \rho^2}\right)^i, \frac{2\rho^2\sigma^2 + \rho^4}{b(\rho^2 + \sigma^2)}\sum_{i=0}^{t-1}\left(\frac{\sigma^4}{(\sigma^2 + \rho^2)^2}\right)^i\right).$$

Splitting strategy 2 – Similar calculus as in the above section can be undertaken by simply replacing ρ^2 by $\rho^2 b$.

Appendices – Chapter 5

Chapter contents

D.1 Proof of Proposition 10

(i) Let $\tau, \rho > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. Since f is assumed to be proper, it is lower-bounded and there exists a scalar constant $C \in \mathbb{R}$ such that $f(\mathbf{z}) \geq C$, for all $\mathbf{z} \in \mathbb{R}^d$. It follows that

$$\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \geq C - \tau \log \int_{\mathbb{R}^d} \frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{(2\pi\rho^2)^{d/2}} \mathrm{d}\mathbf{z} > -\infty. \tag{D.1}$$

By definition and since $\mu(\mathrm{dom} f) > 0$, that there exists $M \in \mathbb{R}$ such that for all $\mathbf{z} \in \mathrm{dom} f$, $f(\mathbf{z}) \leq M$. This yields:

$$\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \leq M - \tau \log \int_{\mathrm{dom} f} \frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{(2\pi\rho^2)^{d/2}} \mathrm{d}\mathbf{z} < \infty. \tag{D.2}$$

Therefore, $\mathcal{A}_\rho^\tau[f]$ is proper and its domain is $\mathbb{R}^d$.

(ii) & (iii) These properties follow from Fatou's lemma, the dominated convergence theorem and the continuity of $\mathbf{x} \mapsto \exp(-\|\mathbf{x}\|^2)$.

(iv) See property (iii) of Proposition 1 in Chapter 1.

D.2 Proof of Proposition 11

From property (iv) of Proposition 10, $\mathcal{A}_\rho^\tau[f]$ is strictly convex and hence admits a unique minimum. Let $\pi \propto \exp(-f)$ be the normalized probability density function associated to the potential function f. Since f is symmetric around $\boldsymbol{\theta}_0$, it follows that $\mathbb{E}_\pi(\vartheta) = \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \pi(\boldsymbol{\theta}) = \boldsymbol{\theta}_0$. Then, the result follows from property (i) of Proposition 1 in Chapter 1.

D.3 Proof of Proposition 12

(i) We first show that when f is Lipschitz continuous, then so is $\mathcal{A}_\rho^\tau[f]$.

Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, $\tau > 0$ and $\rho > 0$. To begin with, we assume that

$$\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u} \geq \int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}.$$

Then, we have

$$
\begin{aligned}
\left|\mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}_2) - \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}_1)\right| &= \tau \left| \log \frac{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}} \right| \\
&= \tau \log \frac{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}} \\
&= \tau \log \int_{\mathbb{R}^d} \frac{\exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right)}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}} \\
&\quad \times \exp\left(\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u})}{\tau}\right) \mathrm{d}\mathbf{u} \\
&\leq \tau \log \max_{\mathbf{u}\in\mathbb{R}^d} \left\{ \exp\left(\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u})}{\tau}\right) \right\} \\
&\quad \times \frac{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}} \\
&\leq \max_{\mathbf{u}\in\mathbb{R}^d} \left\{ f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u}) - f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u}) \right\} \\
&\leq L \left\| \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \right\|.
\end{aligned}
$$

With similar calculations, the case

$$
\int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_1 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u} \leq \int_{\mathbb{R}^d} \exp\left(-\frac{f(\boldsymbol{\theta}_2 + \rho\sqrt{\tau}\mathbf{u})}{\tau} - \frac{\|\mathbf{u}\|^2}{2}\right) \mathrm{d}\mathbf{u}
$$

leads to the same result.

(ii) We now show the gradient-Lipschitz property of $\mathcal{A}_\rho^\tau[f]$. From property (iv) in Proposition 1, we know that $\mathcal{A}_\rho^\tau[f]$ is continuously differentiable. Let denote

$$
A(\boldsymbol{\theta}, \mathbf{z}) = \exp\left(-\frac{f(\mathbf{z})}{\tau} - \frac{\|\boldsymbol{\theta} - \mathbf{z}\|^2}{2\rho^2\tau}\right).
$$

Its gradient and its Hessian are defined for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by

$$\nabla \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) = \frac{1}{\rho^2} \cdot \frac{\displaystyle\int_{\mathbb{R}^d} (\boldsymbol{\theta} - \mathbf{z}) A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z}}{\displaystyle\int_{\mathbb{R}^d} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z}}$$

$$\nabla^2 \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) = \frac{1}{\rho^2} \mathbf{I}_d + \frac{\displaystyle\int_{\mathbb{R}^d} \mathbf{z} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \left(\int_{\mathbb{R}^d} \mathbf{z} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \right)^T - \int_{\mathbb{R}^d} \mathbf{z}\mathbf{z}^T A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \int_{\mathbb{R}^d} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z}}{\rho^2 \tau \left(\displaystyle\int_{\mathbb{R}^d} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \right)^2}.$$

By applying a matrix generalization of the Cauchy-Schwarz inequality (Tripathi, 1999), it follows that

$$\int_{\mathbb{R}^d} \mathbf{z} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \left(\int_{\mathbb{R}^d} \mathbf{z} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \right)^T \preccurlyeq \int_{\mathbb{R}^d} \mathbf{z}\mathbf{z}^T A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z} \int_{\mathbb{R}^d} A(\boldsymbol{\theta}, \mathbf{z}) \mathrm{d}\mathbf{z}$$

Hence, we have

$$\nabla^2 \mathcal{A}_\rho^\tau[f](\boldsymbol{\theta}) \preccurlyeq \frac{1}{\rho^2} \mathbf{I}_d,$$

which shows that $\nabla \mathcal{A}_\rho^\tau[f]$ is $1/\rho^2$-Lipschitz.

D.4 Proof of Proposition 13

(i) The Jensen inequality applied to the convex function $x \mapsto \exp(-x)$ yields

$$\int_{\mathbb{R}^d} \exp\left(-\frac{f(\mathbf{z})}{\tau} \right) \frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2\tau} \right)}{(2\pi\rho^2\tau)^{d/2}} \mathrm{d}\mathbf{z} \geq \exp\left(-\int_{\mathbb{R}^d} \frac{f(\mathbf{z})}{\tau} \frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2\tau} \right)}{(2\pi\rho^2\tau)^{d/2}} \mathrm{d}\mathbf{z} \right),$$

which gives the desired result by taking the natural logarithm and multiplying by $-\tau$ on both sides.
(ii) Let $\boldsymbol{\theta} \in \mathbb{R}^d$ and

$$\mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta}) \coloneqq \operatorname*{arg\,min}_{\mathbf{z} \in \mathbb{R}^d} \left\{ f(\mathbf{z}) + \frac{\|\mathbf{z} - \boldsymbol{\theta}\|^2}{2\rho^2} \right\}.$$

Since $\rho^{-2}(\boldsymbol{\theta} - \mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta})) \in \partial f(\mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta}))$, the convexity of f yields for all $\mathbf{z} \in \mathbb{R}^d$,

$$f(\mathbf{z}) \geq f(\mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta})) + \langle \rho^{-2}(\boldsymbol{\theta} - \mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta})), \mathbf{z} - \mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta}) \rangle.$$

This implies that

$$
\begin{aligned}
&\int_{\mathbb{R}^d} \exp\left(-\frac{f(\mathbf{z})}{\tau}\right) \frac{\exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2\tau}\right)}{(2\pi\rho^2\tau)^{d/2}} \mathrm{d}\mathbf{z} \\
&\leq \exp\left(-\frac{f(\mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta}))}{\tau}\right) \int_{\mathbb{R}^d} \frac{\exp\left(-\frac{1}{\rho^2\tau}\langle \boldsymbol{\theta} - \mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta}), \mathbf{z} - \mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta})\rangle - \frac{\|\mathbf{z}-\boldsymbol{\theta}\|^2}{2\rho^2\tau}\right)}{(2\pi\rho^2\tau)^{d/2}} \mathrm{d}\mathbf{z} \\
&= \exp\left(-\frac{f(\mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta}))}{\tau} - \frac{\left\|\boldsymbol{\theta} - \mathrm{prox}_{\rho^2 f}(\boldsymbol{\theta})\right\|^2}{2\rho^2\tau}\right) \\
&= \exp\left(-\frac{\mathcal{M}_\rho[f](\boldsymbol{\theta})}{\tau}\right).
\end{aligned}
$$

Then, taking the natural logarithm and multiplying by $-\tau$ on both sides gives the desired result.

Bibliography

Adler, S. L. (1981). "Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions." *Physical Review D* 23 (12): 2901–2904. doi:10.1103/PhysRevD.23.2901. (Cited on page 101).

Afonso, M. V., J. M. Bioucas-Dias, and M. A. T. Figueiredo (2010). "Fast Image Recovery Using Variable Splitting and Constrained Optimization." *IEEE Transactions on Image Processing* 19 (9): 2345–2356. doi:10.1109/TIP.2010.2047910. (Cited on pages 56, 69).

Albert, J. H., and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88 (422): 669–679. doi:10.2307/2290350. (Cited on page 33).

Altmann, Y., S. McLaughlin, and N. Dobigeon (2014). "Sampling from a multivariate Gaussian distribution truncated on a simplex: a review." In *Proc. IEEE Workshop on Statistical Signal Processing (SSP),* 113–116. Invited paper. Gold Coast, Australia, July. doi:10.1109/SSP.2014.6884588. (Cited on pages 44, 98).

Ambrosio, L., N. Gigli, and G. Savaré (2008). *Gradient flows in metric spaces and in the space of probability measures.* 2nd. Lectures in Mathematics. ETH Zürich. Available at https://www.springer.com/gp/book/9783764373092. Birkhäuser Verlag. (Cited on page 135).

Andrieu, C., N. de Freitas, A. Doucet, and M. I. Jordan (2003). "An Introduction to MCMC for Machine Learning." *Machine Learning* 50 (1–2): 5–43. doi:10.1023/A:1020281327116. (Cited on page 18).

Andrieu, C., and É. Moulines (2006). "On the ergodicity properties of some adaptive MCMC algorithms." *Annals of Applied Probability* 16, no. 3 (August): 1462–1505. doi:10.1214/105051606000000286. (Cited on page 19).

Andrieu, C., and C. P. Robert (2001). *Controlled MCMC for Optimal Sampling.* [online]. Technical report. Available at http://crest.science/RePEc/wpstorage/2001-33.pdf. (Cited on page 19).

Anscombe, F. J. (1948). "The transformation of Poisson, binomial and negative-binomial data." *Biometrika* 35:246–254. doi:10.2307/2332343. (Cited on page 65).

Armstrong, S. A., J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer (2002). "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia." *Nature Genetics* 30 (1): 41–47. doi:10.1038/ng765. (Cited on page 100).

Arnold, L. (1974). *Stochastic Differential Equations: Theory and Applications*. Wiley-Interscience. (Cited on page 152).

Arrow, K. J., L. Hurwicz, and H. Uzawa (1958). *Studies in linear and non-linear programming*. Stanford University Press. (Cited on page 112).

Azoury, K. S., and M. K. Warmuth (2001). "Relative Loss Bounds for On-Line Density Estimation with the Exponential Family of Distributions." *Machine Learning* 43, no. 3 (June): 211–246. doi:10.1023/A:1010896012157. (Cited on page 32).

Baldassi, C., C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina (2016). "Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes." *Proceedings of the National Academy of Sciences* 113 (48): E7655–E7662. doi:10.1073/pnas.1608103113. (Cited on page 119).

Baldassi, C., A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina (2015). "Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses." *Phys. Rev. Lett.* 115 (12): 128101. doi:10.1103/PhysRevLett.115.128101. (Cited on page 119).

Banerjee, A., S. Merugu, I. S. Dhillon, and J. Ghosh (2005). "Clustering with Bregman Divergences." Available at http://www.jmlr.org/papers/volume6/banerjee05b/banerjee05b.pdf, *Journal of Machine Learning Research* 6:1705–1749. (Cited on page 32).

Baragatti, M., A. Grimaud, and D. Pommeret (2013). "Likelihood-free parallel tempering." *Statistics and Computing* 23 (4): 535–549. doi:10.1007/s11222-012-9328-6. (Cited on page 37).

Barbos, A.-C., F. Caron, J.-F. Giovannelli, and A. Doucet (2017). "Clone MCMC: Parallel High-Dimensional Gaussian Gibbs Sampling." In *Advances in Neural Information Processing Systems*, 5020–5028. Available at https://papers.nips.cc/paper/7087-clone-mcmc-parallel-high-dimensional-gaussian-gibbs-sampling. (Cited on pages 34, 48, 102, 103, 104, 110).

Bardenet, R., A. Doucet, and C. Holmes (2017). "On Markov chain Monte Carlo methods for tall data." Available at http://www.jmlr.org/papers/volume18/15-205/15-205.pdf, *Journal of Machine Learning Research* 18 (47): 1–43. (Cited on pages 20, 21).

Barone, P., and A. Frigessi (1990). "Improving Stochastic Relaxation for Gaussian Random Fields." *Probability in the Engineering and Informational Sciences* 4 (3): 369–389. doi:10.1017/S0269964800001674. (Cited on page 101).

Barron, A. R., and T. M. Cover (1991). "Minimum complexity density estimation." *IEEE Transactions on Information Theory* 37 (4): 1034–1054. doi:10.1109/18.86996. (Cited on page 73).

Bauschke, H. H., and P. L. (Combettes (2013). *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. (Cited on page 111).

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). "Approximate Bayesian Computation in Population Genetics." Available at https://www.genetics.org/content/genetics/162/4/2025.full.pdf, *Genetics* 162 (4): 2025–2035. (Cited on page 37).

Beck, A. (2015). "On the Convergence of Alternating Minimization for Convex Programming with Applications to Iteratively Reweighted Least Squares and Decomposition Schemes." *SIAM Journal on Optimization* 25:185–209. doi:10.1137/13094829X. (Cited on page 56).

Beck, A. (2017). *First-Order Methods in Optimization.* Society for Industrial / Applied Mathematics. doi:10.1137/1.9781611974997. (Cited on pages 80, 123).

Beck, A., and M. Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization." *Operations Research Letters* 31 (3): 167–175. doi:10.1016/S0167-6377(02)00231-6. (Cited on page 31).

Beck, A., and M. Teboulle (2012). "Smoothing and First Order Methods: A Unified Framework." *SIAM Journal on Optimization* 22 (2): 557–580. doi:10.1137/100818327. (Cited on pages 118, 119, 123, 124).

Bellman, R., R. E. Kalaba, and J. A. Lockett (1966). *Numerical inversion of the Laplace transform.* Elsevier. (Cited on page 112).

Ben-Tal, A., T. Margalit, and A. Nemirovski (2001). "The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography." *SIAM Journal on Optimization* 12, no. 1 (January): 79–108. doi:10.1137/S1052623499354564. (Cited on page 31).

Bennett, J., and N. Bez (2015). "Generating monotone quantities for the heat equation." *Journal für die reine und angewandte Mathematik (Crelles Journal).* doi:10.1515/crelle-2017-0025. (Cited on page 146).

Besag, J., and C. Kooperberg (1995). "On conditional and intrinsic autoregressions." *Biometrika* 82 (4): 733–746. doi:10.2307/2337341. (Cited on page 98).

Besag, J. (1986). "On the Statistical Analysis of Dirty Pictures." *Journal of the Royal Statistical Society: Series B (Methodological)* 48 (3): 259–279. doi:10.1111/j.2517-6161.1986.tb01412.x. (Cited on page 57).

Betancourt, M. (2011). "Nested Sampling with Constrained Hamiltonian Monte Carlo." *AIP Conference Proceedings* 1305 (1): 165–172. doi:10.1063/1.3573613. (Cited on page 44).

Bioucas-Dias, J. M., and M. A. T. Figueiredo (2007). "A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration." *IEEE Transactions on Image Processing* 16 (12): 2992–3004. doi:10.1109/TIP.2007.909319. (Cited on page 21).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer-Verlag. (Cited on page 99).

Bobin, J., J.-L. Starck, Y. Moudden, and M. J. Fadili (2008). "Chapter 5 - Blind Source Separation: The Sparsity Revolution," edited by P. W. Hawkes, 152:221–302. Advances in Imaging and Electron Physics. Elsevier. doi:10.1016/S1076-5670(08)00605-8. (Cited on page 133).

Box, G. E. P., and M. E. Muller (1958). "A Note on the Generation of Random Normal Deviates." *Annals of Mathematical Statistics* 29, no. 2 (June): 610–611. doi:10.1214/aoms/1177706645. (Cited on page 18).

Box, G. E. P., and G. M. Jenkins (1994). *Time Series Analysis: Forecasting and Control.* 3rd. Prentice Hall PTR. (Cited on page 108).

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers." Available at https://web.stanford.edu/\protect\unhbox\voidb@x\penalty\@M\{}boyd/papers/pdf/admm_distr_stats.pdf, *Foundations and Trends in Machine Learning* 3 (1): 1–122. (Cited on pages 39, 113).

Boyd, S., and L. Vandenberghe (2004). *Convex Optimization.* 7th. Cambridge University Press. (Cited on page 28).

Bramble, J., J. Pasciak, and A. Vassilev (1997). "Analysis of the Inexact Uzawa Algorithm for Saddle Point Problems." *SIAM Journal on Numerical Analysis* 34 (3): 1072–1092. doi:10.1137/S0036142994273343. (Cited on page 112).

Bredies, K., K. Kunisch, and T. Pock (2010). "Total Generalized Variation." *SIAM Journal on Imaging Sciences* 3 (3): 492–526. doi:10.1137/090769521. (Cited on page 48).

Bredies, K., and H. Sun (2017). "A Proximal Point Analysis of the Preconditioned Alternating Direction Method of Multipliers." *Journal of Optimization Theory and Applications* 173, no. 3 (June): 878–907. doi:10.1007/s10957-017-1112-5. (Cited on page 113).

Bregman, L. (1967). "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." *USSR Computational Mathematics and Mathematical Physics* 7 (3): 200–217. doi:https://doi.org/10.1016/0041-5553(67)90040-7. (Cited on page 32).

Brosse, N., A. Durmus, É. Moulines, and M. Pereyra (2017). "Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo." In *Conference on Learning Theory,* 65:319–342. Available at http://proceedings.mlr.press/v65/brosse17a.html. (Cited on pages 44, 45, 148).

Brosse, N., É. Moulines, and A. Durmus (2018). "The Promises and Pitfalls of Stochastic Gradient Langevin Dynamics." In *Neural Information Processing Systems,* 8278–8288. Available at https://papers.nips.cc/paper/8048-the-promises-and-pitfalls-of-stochastic-gradient-langevin-dynamics.pdf. Montréal, Canada. (Cited on pages 21, 57).

Bucher, C. G. (1988). "Adaptive sampling — an iterative fast Monte Carlo procedure." *Structural Safety* 5 (2): 119–126. doi:10.1016/0167-4730(88)90020-3. (Cited on page 18).

Candes, E. J., and Y. Plan (2010). "Matrix Completion With Noise." *Proceedings of the IEEE* 98 (6): 925–936. doi:10.1109/JPROC.2009.2035722. (Cited on page 17).

Canny, J. (2004). "GaP: A Factor Model for Discrete Data." In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 122–129. doi:10.1145/1008992.1009016. (Cited on page 34).

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76 (1): 1–32. doi:10.18637/jss.v076.i01. (Cited on pages 20, 68).

Celeux, G., S. Chretien, F. Forbes, and A. Mkhadri (2001). "A Component-Wise EM Algorithm for Mixtures." *Journal of Computational and Graphical Statistics* 10 (4): 697–712. doi:10.1198/106186001317243403. (Cited on page 29).

Celeux, G., M. El Anbari, J.-M. Marin, and C. P. Robert (2012). "Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation." *Bayesian Analysis* 7, no. 2 (June): 477–502. doi:10.1214/12-BA716. (Cited on page 44).

Chambolle, A., M. Novaga, D. Cremers, and T. Pock (2010). "An introduction to total variation for image analysis." In *Theoretical Foundations and Numerical Methods for Sparse Recovery, De Gruyter.* Available at https://hal.archives-ouvertes.fr/hal-00437581/document. (Cited on pages 48, 61).

Chambolle, A. (2004). "An Algorithm for Total Variation Minimization and Applications." *Journal of Mathematical Imaging and Vision* 20, no. 1 (January): 89–97. doi:10.1023/B:JMIV.0000011325.36760.1e. (Cited on pages 61, 70).

Chambolle, A., and T. Pock (2011). "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging." *Journal of Mathematical Imaging and Vision* 40, no. 1 (May): 120–145. doi:10.1007/s10851-010-0251-1. (Cited on pages 113, 132).

Chang, E. S., C. Hung, W. Liu, and J. Yina (2016). "A Denoising algorithm for remote sensing images with impulse noise." In *IEEE International Geoscience and Remote Sensing Symposium,* 2905–2908. doi:10.1109/IGARSS.2016.7729750. (Cited on page 58).

Chaudhari, P., A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina (2019). "Entropy-SGD: biasing gradient descent into wide valleys." *Journal of Statistical Mechanics: Theory and Experiment* 2019, no. 12 (December): 124018. doi:10.1088/1742-5468/ab39d9. (Cited on pages 119, 127).

Chaudhari, P., A. Oberman, S. Osher, S. Soatto, and G. Carlier (2018). "Deep relaxation: partial differential equations for optimizing deep neural networks." *Research in the Mathematical Sciences* 5 (3). doi:10.1007/s40687-018-0148-y. (Cited on page 122).

Chen, Z., and S. S. Vempala (2019). "Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Log-concave Distributions." In *Proceedings of the International Conference on Randomization and Computation*. Available at https://drops.dagstuhl.de/opus/volltexte/2019/11279/pdf/LIPIcs-APPROX-RANDOM-2019-64.pdf. (Cited on pages 20, 84).

Cheng, X., and P. Bartlett (2018). "Convergence of Langevin MCMC in KL-divergence." In *Proceedings of Algorithmic Learning Theory*, 186–211. Available at https://research.cs.cornell.edu/conferences/alt2018/A/cheng18.pdf. (Cited on page 85).

Cheng, X., N. S. Chatterji, P. L. Bartlett, and M. I. Jordan (2018). "Underdamped Langevin MCMC: A non-asymptotic analysis." In *Proceedings of the 31st Conference On Learning Theory*, 300–323. Available at http://proceedings.mlr.press/v75/cheng18a.html. (Cited on pages 20, 84).

Choi, H. M., and J. P. Hobert (2013). "The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic." *Electronic Journal of Statistics* 7:2054–2064. doi:doi:10.1214/13-EJS837. (Cited on pages 29, 79).

Combettes, P. L., and J.-C. Pesquet (2011). "Proximal Splitting Methods in Signal Processing." In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, edited by H. H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz, 185–212. Available at https://www.ljll.math.upmc.fr/\protect\unhbox\voidb@x\penalty\@M\{}plc/prox.pdf. Springer. (Cited on pages 21, 57).

Dai Pra, P., B. Scoppola, and E. Scoppola (2012). "Sampling from a Gibbs measure with pair interaction by means of PCA." *Journal of Statistical Physics* 149 (4): 722–737. doi:10.1007/s10955-012-0612-9. (Cited on page 55).

Dalalyan, A. S., and L. Riou-Durand (2018). "On sampling from a log-concave density using kinetic Langevin diffusions." *Bernoulli* 26 (3): 1956–1988. doi:doi:10.3150/19-BEJ1178. (Cited on page 84).

Dalalyan, A. S. (2017). "Theoretical guarantees for approximate sampling from smooth and log-concave densities." *Journal of the Royal Statistical Society, Series B* 79 (3): 651–676. doi:10.1111/rssb.12183. (Cited on pages 20, 80, 85, 86, 88, 90).

Dalalyan, A. S., and A. Karagulyan (2019). "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." *Stochastic Processes and Their Applications* 129 (12): 5278–5311. doi:10.1016/j.spa.2019.02.016. (Cited on page 79).

Dang, Q. A., and M. Ehrhardt (2012). "On Dirac delta sequences and their generating functions." *Applied Mathematics Letters* 25 (12): 2385–2390. doi:10.1016/j.aml.2012.07.009. (Cited on page 31).

Darbon, J., and G. P. Langlois (2020). "On Bayesian posterior mean estimators in imaging sciences and Hamilton-Jacobi Partial Differential Equations." arXiv: 2003.05572. (Cited on pages 121, 122, 124, 125, 127, 128, 132).

Del Moral, P., A. Doucet, and A. Jasra (2012). "An adaptive sequential Monte Carlo method for approximate Bayesian computation." *Statistics and Computing* 22 (5): 1009–1020. doi:10.1007/s11222-011-9271-y. (Cited on page 37).

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B* 39 (1): 1–38. doi:10.1111/j.2517-6161.1977.tb01600.x. (Cited on page 29).

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag. (Cited on page 18).

Dharmadhikari, S., and K. Joag-Dev (1988). *Unimodality, Convexity, and Applications*. New York: Academic Press. (Cited on page 135).

Dietrich, C. R., and G. N. Newsam (1997). "Fast and Exact Simulation of Stationary Gaussian Processes through Circulant Embedding of the Covariance Matrix." *SIAM Journal on Scientific Computing* 18 (4): 1088–1107. doi:10.1137/S1064827592240555. (Cited on page 97).

Dobson, A. J., and A. G. Barnett (2008). *An Introduction to Generalized Linear Models*. 3rd ed. Texts in Statistical Science. Boca Raton, FL: Chapman & Hall/CRC Press. (Cited on page 34).

Dolcetta, I. C. (2003). "Representations of Solutions of Hamilton-Jacobi Equations." In *Nonlinear Equations: Methods, Models and Applications*, edited by D. Lupo, C. D. Pagani, and B. Ruf, 79–90. Available at https://link.springer.com/chapter/10.1007/978-3-0348-8087-9_6. (Cited on page 122).

Doucet, A., S. J. Godsill, and C. P. Robert (2002). "Marginal maximum a posteriori estimation using Markov chain Monte Carlo." *Statistics and Computing* 12 (1): 77–84. doi:10.1023/A:1013172322619. (Cited on page 29).

Duane, S., A. Kennedy, B. J. Pendleton, and D. Roweth (1987). "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–222. doi:10.1016/0370-2693(87)91197-X. (Cited on pages 19, 29, 53, 57, 65).

Duchi, J. C., A. Agarwal, M. Johansson, and M. I. Jordan (2012). "Ergodic Mirror Descent." *SIAM Journal on Optimization* 22 (4): 1549–1578. doi:10.1137/110836043. (Cited on page 31).

Dümbgen, L., and K. Rufibach (2009). "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency." *Bernoulli* 15 (1): 40–68. doi:10.3150/08-BEJ141. (Cited on page 39).

Dupé, F., J. M. Fadili, and J. Starck (2009). "A Proximal Iteration for Deconvolving Poisson Noisy Images Using Sparse Representations." *IEEE Transactions on Image Processing* 18 (2): 310–321. doi:10.1109/TIP.2008.2008223. (Cited on pages 54, 65, 73).

Dupé, F., J. M. Fadili, and J. Starck (2012). "Deconvolution under Poisson noise using exact data fidelity and synthesis or analysis sparsity priors." *Statistical Methodology* 9 (1): 4–18. doi:10.1016/j.stamet.2011.04.008. (Cited on page 65).

Durmus, A., E. Moulines, and M. Pereyra (2018). "Efficient Bayesian Computation by Proximal Markov chain Monte Carlo: When Langevin Meets Moreau." *SIAM Journal on Imaging Sciences* 11 (1): 473–506. doi:10.1137/16M1108340. (Cited on pages 17, 21, 57, 61, 63, 65, 70, 155).

Durmus, A., S. Majewski, and B. Miasojedow (2019). "Analysis of Langevin Monte Carlo via convex optimization." Available at http://www.jmlr.org/papers/volume20/18-173/18-173.pdf, *Journal of Machine Learning Research* 20 (73): 1–46. (Cited on pages 80, 84, 85).

Durmus, A., and E. Moulines (2017). "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm." *The Annals of Applied Probability* 27, no. 3 (June): 1551–1587. doi:10.1214/16-AAP1238. (Cited on pages 20, 79, 85).

Durmus, A., and E. Moulines (2019). "High-dimensional Bayesian inference via the unadjusted Langevin algorithm." *Bernoulli* 25 (4A): 2854–2882. doi:10.3150/18-BEJ1073. (Cited on pages 83, 162).

Durmus, A., E. Moulines, and S. Eero (2017). "On the convergence of Hamiltonian Monte Carlo." arXiv: 1705.00166. (Cited on page 20).

Dwivedi, R., Y. Chen, M. J. Wainwright, and B. Yu (2019). "Log-concave sampling: Metropolis-Hastings algorithms are fast." Available at http://www.jmlr.org/papers/volume20/19-306/19-306.pdf, *Journal of Machine Learning Research* 20 (183): 1–42. (Cited on pages 20, 85, 86, 88).

Van Dyk, D. A., and X.-L. Meng (2001). "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics* 10 (1): 1–50. doi:10.1198/10618600152418584. (Cited on pages 23, 27, 29).

Eckstein, J., and D. P. Bertsekas (1992). "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators." *Mathematical Programming* 55 (1): 293–318. doi:10.1007/BF01581204. (Cited on page 65).

Edwards, R. G., and A. D. Sokal (1988). "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm." *Physical Review D* 38 (6): 2009–2012. doi:10.1103/PhysRevD.38.2009. (Cited on page 29).

Efron, B. (2011). "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association* 106 (496): 1602–1614. doi:10.1198/jasa.2011.tm11181. (Cited on page 120).

Elad, M. (2006). "Why Simple Shrinkage Is Still Relevant for Redundant Representations?" *IEEE Transactions on Information Theory* 52 (12): 5559–5569. doi:10.1109/TIT.2006.885522. (Cited on page 21).

Esser, E., X. Zhang, and T. Chan (2010). "A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science." *SIAM Journal on Imaging Sciences* 3 (4): 1015–1046. doi:10.1137/09076934X. (Cited on page 113).

Fadili, J., and J.-L. Starck (2005). "EM algorithm for sparse representation-based image inpainting." In *IEEE International Conference on Image Processing*, 2:II–61. doi:10.1109/ICIP.2005.1529991. (Cited on page 116).

Fadili, M., and E. Bullmore (2002). "Wavelet-Generalized Least Squares: A New BLU Estimator of Linear Regression Models with 1/f Errors." *NeuroImage* 15 (1): 217–232. doi:10.1006/nimg.2001.0955. (Cited on page 116).

Fearnhead, P., and D. Prangle (2012). "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation." *Journal of the Royal Statistical Society, Series B* 74 (3): 419–474. doi:10.1111/j.1467-9868.2011.01010.x. (Cited on pages 31, 36).

Fellows, M., A. Mahajan, T. G. J. Rudner, and S. Whiteson (2019). "VIREL: A Variational Inference Framework for Reinforcement Learning." In *Advances in Neural Information Processing Systems*, 7120–7134. Available at https://papers.nips.cc/paper/8934-virel-a-variational-inference-framework-for-reinforcement-learning. (Cited on page 32).

Févotte, C., N. Bertin, and J.-L. Durrieu (2009). "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis." *Neural Computation* 21 (3): 793–830. doi:10.1162/neco.2008.04-08-771. (Cited on page 32).

Figueiredo, M. A. T., and J. M. Bioucas-Dias (2010). "Restoration of Poissonian images using alternating direction optimization." *IEEE Transactions on Image Processing* 19 (12): 3133–3145. doi:10.1109/TIP.2010.2053941. (Cited on pages 64, 65).

Filstroff, L., A. Lumbreras, and C. Févotte (2018). "Closed-form Marginal Likelihood in Gamma-Poisson Matrix Factorization." In *International Conference on Machine Learning*. Available at http://proceedings.mlr.press/v80/filstroff18a.html. (Cited on pages 28, 116).

Folland, G. (1999). *Real Analysis: Modern Techniques and their Applications*. 2nd ed. New York: Wiley. (Cited on page 135).

Fort, G., and E. Moulines (2003). "Polynomial ergodicity of Markov transition kernels." *Stochastic Processes and their Applications* 103 (1): 57–99. doi:10.1016/S0304-4149(02)00182-5. (Cited on page 79).

Fox, C., and A. Parker (2017). "Accelerated Gibbs sampling of normal distributions using matrix splittings and polynomials." *Bernoulli* 23, no. 4B (November): 3711–3743. doi:10.3150/16-BEJ863. (Cited on pages 22, 102, 108, 110).

Gabay, D., and B. Mercier (1976). "A dual algorithm for the solution of nonlinear variational problems via finite element approximation." *Computers & Mathematics with Applications* 2 (1): 17–40. doi:10.1016/0898-1221(76)90003-1. (Cited on page 113).

Van de Geer, S. (2016). *Estimation and Testing Under Sparsity*. 1st ed. Lecture Notes in Mathematics 2159. Springer. (Cited on page 50).

Gelman, A., J. B. Carlin, H. S. Stern, A. Vehtari, and D. B. Rubin (2003). *Bayesian Data Analysis*. 2nd. Chapman / Hall/CRC. (Cited on pages 17, 101).

Geman, D., and C. Yang (1995). "Nonlinear image recovery with half-quadratic regularization." *IEEE Transactions on Image Processing* 4 (7): 932–946. doi:10.1109/83.392335. (Cited on page 29).

Geman, S., and D. Geman (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (6): 721–741. doi:10.1109/TPAMI.1984.4767596. (Cited on pages 18, 53, 99, 101).

Geweke, J. (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57 (6): 1317–1339. doi:10.2307/1913710. (Cited on page 18).

Gilavert, C., S. Moussaoui, and J. Idier (2015). "Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems Using MCMC." *IEEE Transactions on Signal Processing* 63, no. 1 (January): 70–80. doi:10.1109/TSP.2014.2367457. (Cited on pages 22, 67, 68).

Gilks, W. R., and P. Wild (1992). "Adaptive Rejection Sampling for Gibbs Sampling." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41 (2): 337–348. doi:10.2307/2347565. (Cited on pages 18, 35).

Gilks, W., S. Richardson, and D. Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice*. 1st ed. Chapman & Hall. Taylor & Francis. (Cited on page 18).

Gilks, W. R., G. O. Roberts, and S. K. Sahu (1998). "Adaptive Markov Chain Monte Carlo through Regeneration." *Journal of the American Statistical Association* 93 (443): 1045–1054. doi:10.1080/01621459.1998.10473766. (Cited on page 19).

Giovannelli, J. F. (2008). "Unsupervised Bayesian Convex Deconvolution Based on a Field With an Explicit Partition Function." *IEEE Transactions on Image Processing* 17 (1): 16–26. doi:10.1109/TIP.2007.911819. (Cited on page 39).

Girolami, M., and B. Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (2): 123–214. doi:10.1111/j.1467-9868.2010.00765.x. (Cited on page 20).

Givens, C. R., and R. M. Shortt (1984). "A class of Wasserstein metrics for probability distributions." *Michigan Mathematical Journal* 31 (2): 231–240. doi:10.1307/mmj/1029003026. (Cited on pages 155, 157, 159).

Glowinski, R., and A. Marroco (1975). "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires." Available at http://www.numdam.org/item/M2AN_1975__9_2_41_0, *R.A.I.R.O. Analyse Numérique* 9:41–76. (Cited on page 113).

Golub, G. H., and C. F. Van Loan (1989). *Matrix Computations.* 2nd. The Johns Hopkins University Press. (Cited on page 102).

Goodfellow, I. J., Y. Bengio, and A. Courville (2016). *Deep Learning.* Cambridge, MA, USA: MIT Press. (Cited on page 27).

Goodman, J., and A. D. Sokal (1989). "Multigrid Monte Carlo method. Conceptual foundations." *Physical Review D* 40 (6): 2035–2071. doi:10.1103/PhysRevD.40.2035. (Cited on page 101).

Gradshteyn, I. S., and I. M. Ryzhik (2015). *Table of Integrals, Series, and Products.* 8th ed. London: Academic Press. (Cited on pages 138, 140, 141).

Gu, S., L. Zhang, W. Zuo, and X. Feng (2014). "Weighted Nuclear Norm Minimization with Application to Image Denoising." In *IEEE Conference on Computer Vision and Pattern Recognition,* 2862–2869. doi:10.1109/CVPR.2014.366. (Cited on page 81).

Haario, H., E. Saksman, and J. Tamminen (2001). "An adaptive Metropolis algorithm." Available at https://projecteuclid.org/download/pdf_1/euclid.bj/1080222083, *Bernoulli* 7, no. 2 (April): 223–242. (Cited on page 19).

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57, no. 1 (April): 97–109. doi:10.1093/biomet/57.1.97. (Cited on page 18).

Higdon, D. M. (1998). "Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications." *Journal of the American Statistical Association* 93 (442): 585–595. doi:10.1080/01621459.1998.10473712. (Cited on page 29).

Higdon, D. (2007). "A Primer on Space-Time Modeling from a Bayesian Perspective." In *Statistical Methods for Spatio-Temporal Systems,* edited by L. H. B. Finkenstadt and V. Isham, 217–279. Chapman & Hall/CRC. (Cited on page 46).

Holmes, C. C., and B. K. Mallick (2003). "Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines." *Journal of the American Statistical Association* 98 (462): 352–368. doi:10.1198/016214503000143. (Cited on pages 32, 33, 34, 39).

Hunter, D. R., and K. Lange (2004). "A Tutorial on MM Algorithms." *The American Statistician* 58 (1): 30–37. doi:10.1198/0003130042836. (Cited on page 21).

Idier, J., ed. (2008). *Bayesian Approach to Inverse Problems.* Wiley. doi:10.1002/9780470611197. (Cited on pages 17, 58, 96, 104, 116).

Ivanciuc, O. (2006). *Comparative Evaluation of Prediction Algorithms.* Available at http://www.coepra.org/CoEPrA_regr.html. (Cited on page 100).

Jarner, S. F., and E. Hansen (2000). "Geometric ergodicity of Metropolis algorithms." *Stochastic Processes and their Applications* 85 (2): 341–361. doi:10.1016/S0304-4149(99)00082-4. (Cited on pages 19, 79).

Jeffreys, H. (1946). "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186 (1007): 453–461. doi:10.1098/rspa.1946.0056. (Cited on page 99).

Johnson, M., J. Saunderson, and A. Willsky (2013). "Analyzing Hogwild Parallel Gaussian Gibbs Sampling." In *Neural Information Processing Systems,* 2715–2723. Available at https://papers.nips.cc/paper/5043-analyzing-hogwild-parallel-gaussian-gibbs-sampling. (Cited on pages 102, 103, 110).

Johnson, V. E., and J. H. Albert (2006). *Ordinal Data Modeling.* Springer Science & Business Media. (Cited on page 44).

Jones, G. L., and J. P. Hobert (2001). "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo." *Statistical Science* 16, no. 4 (November): 312–334. doi:10.1214/ss/1015346317. (Cited on page 79).

Kallio, M., and A. Ruszczynski (1994). *Perturbation Methods for Saddle Point Computation.* Available at http://pure.iiasa.ac.at/id/eprint/4171/1/WP-94-038.pdf. International Institute for Applied Systems Analysis, May. (Cited on page 112).

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). "Markov Chain Monte Carlo in Practice: A Roundtable Discussion." *The American Statistician* 52 (2): 93–100. doi:10.1080/00031305.1998.10480547. (Cited on page 68).

Klein, J. P., and M. L. Moeschberger (2005). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer Science & Business Media. (Cited on page 44).

Krichene, W., A. Bayen, and P. L. Bartlett (2015). "Accelerated Mirror Descent in Continuous and Discrete Time." In *Advances in Neural Information Processing Systems,* 2845–2853. Available at https://papers.nips.cc/paper/5843-accelerated-mirror-descent-in-continuous-and-discrete-time. (Cited on page 31).

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images.* Available at http://www.cs.toronto.edu/~kriz/cifar.html. (Cited on page 27).

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis* 5, no. 2 (June): 369–411. doi:10.1214/10-BA607. (Cited on page 62).

Lawler, G. F. (2010). *Random walk and the heat equation.* Vol. 55. American Mathematical Society. (Cited on page 146).

Le Cun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86 (11): 2278–2324. doi:10.1109/5.726791. (Cited on page 100).

Van Leeuwen, T., and F. J. Herrmann (2015). "A penalty method for PDE-constrained optimization in inverse problems." *Inverse Problems* 32, no. 1 (December): 015007. doi:10.1088/0266-5611/32/1/015007. (Cited on page 56).

Lemaréchal, C., and C. Sagastizábal (1997). "Practical Aspects of the Moreau-Yosida Regularization: Theoretical Preliminaries." *SIAM Journal on Optimization* 7 (2): 367–385. doi:10.1137/S1052623494267127. (Cited on page 155).

Lévy, P. (1940). "Sur certains processus stochastiques homogènes." Available at http://www.numdam.org/article/CM_1940__7__283_0.pdf, *Compositio Mathematica* 7:283–339. (Cited on page 153).

Li, M., D. Sun, and K. Toh (2016). "A Majorized ADMM with Indefinite Proximal Terms for Linearly Constrained Convex Composite Optimization." *SIAM Journal on Optimization* 26 (2): 922–950. doi:10.1137/140999025. (Cited on page 113).

Li, Q., and N. Lin (2010). "The Bayesian elastic net." *Bayesian Analysis* 5, no. 1 (March): 151–170. doi:10.1214/10-BA506. (Cited on page 81).

Li, S. Z. (2009). *Markov Random Field Modeling in Image Analysis*. 3rd. Springer. (Cited on page 97).

Li, Y., and S. K. Ghosh (2015). "Efficient Sampling Methods for Truncated Multivariate Normal and Student-t Distributions Subject to Linear Inequality Constraints." *Journal of Statistical Theory and Practice* 9 (4): 712–732. doi:10.1080/15598608.2014.996690. (Cited on page 98).

Liechty, M. W., J. C. Liechty, and P. Müller (2009). "The Shadow Prior." *Journal of Computational and Graphical Statistics* 18 (2): 368–383. doi:10.1198/jcgs.2009.07072. (Cited on pages 32, 34, 39).

Likas, A. C., and N. P. Galatsanos (2004). "A variational approach for Bayesian blind image deconvolution." *IEEE Transactions on Signal Processing* 52 (8): 2222–2233. doi:10.1109/TSP.2004.831119. (Cited on page 58).

Lions, P., and B. Mercier (1979). "Splitting algorithms for the sum of two nonlinear operators." *SIAM Journal on Numerical Analysis* 16 (6): 964–979. doi:10.1137/0716071. (Cited on page 65).

Liu, J. S., W. H. Wong, and A. Kong (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes." *Biometrika* 81 (1): 27–40. doi:10.1093/biomet/81.1.27. (Cited on page 160).

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer. (Cited on page 68).

Liu, J. S., and Y. N. Wu (1999). "Parameter Expansion for Data Augmentation." *Journal of the American Statistical Association* 94 (448): 1264–1274. doi:10.1080/01621459.1999.10473879. (Cited on page 29).

Luu, T., J. Fadili, and C. Chesneau (2020). "Sampling from non-smooth distribution through Langevin diffusion." *Methodology and Computing in Applied Probability (in press)*. HAL: hal-01492056. (Cited on pages 21, 61, 73).

Maddison, C. J., D. Tarlow, and T. Minka (2014). "$A^\star$ Sampling." In *Advances in Neural Information Processing Systems,* 3086–3094. Available at https://papers.nips.cc/paper/5449-a-sampling.pdf. (Cited on pages 118, 126).

Maddison, C. J., D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet (2018). "Hamiltonian Descent Methods." arXiv: 1809.05042. (Cited on page 57).

Mantoglou, A., and J. L. Wilson (1982). "The Turning Bands Method for simulation of random fields using line generation by a spectral method." *Water Resources Research* 18 (5): 1379–1394. doi:10.1029/WR018i005p01379. (Cited on page 97).

Marin, J.-M., P. Pudlo, C. P. Robert, and R. J. Ryder (2012). "Approximate Bayesian computational methods." *Statistics and Computing* 22 (6): 1167–1180. doi:10.1007/s11222-011-9288-2. (Cited on page 23).

Marnissi, Y., D. Abboud, E. Chouzenoux, J.-C. Pesquet, M. El-Badaoui, and A. Benazza-Benyahia (2019). "A Data Augmentation Approach for Sampling Gaussian Models in High Dimension." In *Proceedings of the 27th European Signal Processing Conference.* Coruna, Spain. doi:10.23919/EUSIPCO.2019.8902496. (Cited on pages 104, 105, 106).

Marnissi, Y., E. Chouzenoux, A. Benazza-Benyahia, and J. Pesquet (2020). "Majorize–Minimize Adapted Metropolis–Hastings Algorithm." *IEEE Transactions on Signal Processing* 68:2356–2369. doi:10.1109/TSP.2020.2983150. (Cited on page 21).

Marnissi, Y., E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet (2018). "An Auxiliary Variable Method for Markov Chain Monte Carlo Algorithms in High Dimension." *Entropy* 20 (2). doi:10.3390/e20020110. (Cited on pages 29, 48, 58, 59, 62, 63, 67, 104, 105).

Martinet, B. (1970). "Brève communication. Régularisation d'inéquations variationnelles par approximations successives." *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* 4 (R3): 154–158. (Cited on page 111).

Martinet, B. (1972). "Determination approaché d'un point fixe d'une application pseudocontractante. Cas de l'application prox." *Comptes Rendus de l'Académie des Sciences* (Paris), Série A, 274:163–165. (Cited on page 111).

Mejía, J. M., and I. Rodríguez-Iturbe (1974). "On the synthesis of random field sampling from the spectrum: An application to the generation of hydrologic spatial processes." *Water Resources Research* 10 (4): 705–711. doi:10.1029/WR010i004p00705. (Cited on page 97).

Melville, P., and V. Sindhwani (2010). "Recommender Systems." In *Encyclopedia of Machine Learning,* edited by C. Sammut and G. I. Webb, 829–838. Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_705. (Cited on page 17).

Meng, X.-L., and D. van Dyk (1997). "The EM Algorithm – an Old Folk-song Sung to a Fast New Tune." *Journal of the Royal Statistical Society, Series B* 59 (3): 511–567. doi:10.1111/1467-9868.00082. (Cited on page 29).

Meng, X.-L., and D. van Dyk (1998). "Fast EM-type implementations for mixed effects models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (3): 559–578. doi:10.1111/1467-9868.00140. (Cited on page 29).

Metropolis, N., and S. Ulam (1949). "The Monte Carlo Method." *Journal of the American Statistical Association* 44 (247): 335–341. doi:10.1080/01621459.1949.10483310. (Cited on page 18).

Meyn, S., and R. L. Tweedie (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag. (Cited on page 79).

Molina, R., J. Mateos, and A. K. Katsaggelos (2006). "Blind Deconvolution Using a Variational Approach to Parameter, Image, and Blur Estimation." *IEEE Transactions on Image Processing* 15 (12): 3715–3727. doi:10.1109/TIP.2006.881972. (Cited on page 58).

Molina, R., and B. D. Ripley (1989). "Using spatial models as priors in astronomical image analysis." *Journal of Applied Statistics* 16 (2): 193–206. doi:10.1080/02664768900000017. (Cited on page 58).

Moreau, J. J. (1965). "Proximité et dualité dans un espace hilbertien." *Bulletin de la Société Mathématique de France* 93:273–299. (Cited on pages 21, 44, 57, 61, 65, 112, 115, 124).

Neal, R. M. (2011). "MCMC Using Hamiltonian Dynamics." Chap. 5 in *Handbook of Markov Chain Monte Carlo.* CRC Press. doi:10.1201/b10905-7. (Cited on page 20).

Nesterov, Y., and V. Spokoiny (2017). "Random Gradient-Free Minimization of Convex Functions." *Foundations of Computational Mathematics* 17, no. 2 (April): 527–566. doi:10.1007/s10208-015-9296-2. (Cited on page 119).

Nocedal, J., and S. J. Wright (2006). *Numerical Optimization.* 2nd ed. Springer. (Cited on page 56).

Nummelin, E. (1978). "A splitting technique for Harris recurrent Markov chains." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 43:309–318. (Cited on page 78).

Nummelin, E., and P. Tuominen (1983). "The rate of convergence in Orey's theorem for Harris recurrent Markov chains with applications to renewal theory." *Stochastic Processes and their Applications* 15 (3): 295–311. doi:10.1016/0304-4149(83)90037-6. (Cited on page 78).

Ollivier, Y. (2009). "Ricci curvature of Markov chains on metric spaces." *Journal of Functional Analysis* 256 (3): 810–864. doi:10.1016/j.jfa.2008.11.001. (Cited on pages 80, 160, 163).

Ong, F., P. Milanfar, and P. Getreuer (2019). "Local Kernels That Approximate Bayesian Regularization and Proximal Operators." *IEEE Transactions on Image Processing* 28 (6): 3007–3019. doi:10.1109/TIP.2019.2893071. (Cited on pages 120, 121, 132).

Orieux, F., J.-F. Giovannelli, and T. Rodet (2010). "Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution." *Journal of the Optical Society of America A* 27, no. 7 (July): 1593–1607. doi:10.1364/JOSAA.27.001593. (Cited on pages 97, 104).

Paisley, J., D. M. Blei, and M. I. Jordan (2014). "Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference." In *Handbook of Mixed Membership Models and Their Applications*. CRC Press. doi:10.1201/b17520-15. (Cited on page 44).

Papandreou, G., and A. L. Yuille (2011). "Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models." In *International Conference on Computer Vision,* 193–200. doi:10.1109/ICCV.2011.6126242. (Cited on pages 22, 48, 118).

Park, T., and G. Casella (2008). "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–686. doi:10.1198/016214508000000337. (Cited on pages 47, 48, 65).

Parker, A., and C. Fox (2012). "Sampling Gaussian Distributions in Krylov Spaces with Conjugate Gradients." *SIAM Journal on Scientific Computing* 34 (3): B312–B334. doi:10.1137/110831404. (Cited on pages 22, 98).

Pavliotis, G. A. (2014). *Diffusion Processes, the Fokker-Planck and Langevin equations*. Vol. 60. Texts in Applied Mathematics. Springer, New York. (Cited on pages 152, 153).

Pedemonte, S., C. Catana, and K. Van Leemput (2015). "Bayesian tomographic reconstruction using Riemannian MCMC." In *Med. Image Comp. Computer-Assisted Intervention (MICCAI)*. (Cited on page 65).

Pereira, M., and N. Desassis (2019). "Efficient simulation of Gaussian Markov random fields by Chebyshev polynomial approximation." *Spatial Statistics* 31:100359. doi:10.1016/j.spasta.2019.100359. (Cited on page 22).

Pereyra, M. (2016). "Proximal Markov chain Monte Carlo algorithms." *Statistics and Computing* 26 (4): 745–760. doi:10.1007/s11222-015-9567-4. (Cited on pages 21, 39, 53, 57, 61, 65).

Pereyra, M., P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin (2016). "A Survey of Stochastic Simulation and Optimization Methods in Signal Processing." *IEEE Journal of Selected Topics in Signal Processing* 10 (2): 224–241. doi:10.1109/JSTSP.2015.2496908. (Cited on pages 22, 116).

Pereyra, M. (2019). "Revisiting Maximum-A-Posteriori Estimation in Log-Concave Models." *SIAM Journal on Imaging Sciences* 12 (1): 650–670. doi:10.1137/18M1174076. (Cited on page 72).

Perko, L. (2013). *Differential Equations and Dynamical Systems*. Vol. 7. Springer Science & Business Media. (Cited on page 154).

Polson, N. G., J. G. Scott, and J. Windle (2013). "Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504): 1339–1349. doi:10.1080/01621459.2013.829001. (Cited on pages 29, 33).

Rendell, L. J., A. M. Johansen, A. Lee, and N. Whiteley (2018). *Global consensus Monte Carlo.* [online]. Technical report. Available at https://arxiv.org/abs/1807.09288/. (Cited on pages 34, 36, 37, 51, 55, 131).

Robbins, H., and S. Monro (1951). "A Stochastic Approximation Method." *Annals of Mathematical Statistics* 22, no. 3 (September): 400–407. doi:10.1214/aoms/1177729586. (Cited on page 21).

Robert, C. P. (2001). *The Bayesian Choice: from decision-theoretic foundations to computational implementation.* 2nd ed. New York: Springer. (Cited on pages 17, 42, 64, 99).

Robert, C. P., and G. Casella (2004). *Monte Carlo Statistical Methods.* 2nd ed. Berlin: Springer. (Cited on pages 18, 19, 35, 99, 100, 109, 163).

Roberts, G. O., and S. K. Sahu (1997). "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler." *Journal of the Royal Statistical Society. Series B (Methodological)* 59 (2): 291–317. doi:10.1111/1467-9868.00070. (Cited on page 101).

Roberts, G. O., and O. Stramer (2002). "Langevin Diffusions and Metropolis-Hastings Algorithms." *Methodology And Computing In Applied Probability* 4:337–357. doi:10.1023/A:1023562417138. (Cited on page 79).

Roberts, G. O., and A. F. Smith (1994). "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms." *Stochastic Processes and their Applications* 49 (2): 207–216. doi:10.1016/0304-4149(94)90134-1. (Cited on page 55).

Roberts, G. O., and J. S. Rosenthal (2004). "General state space Markov chains and MCMC algorithms." *Probability Surveys* 1:20–71. doi:10.1214/154957804100000024. (Cited on page 79).

Roberts, G. O., and R. L. Tweedie (1996). "Exponential convergence of Langevin distributions and their discrete approximations." Available at https://projecteuclid.org/download/pdf_1/euclid.bj/1178291835, *Bernoulli* 2, no. 4 (December): 341–363. (Cited on pages 20, 53, 57).

Rockafellar, R. T. (1976). "Monotone Operators and the Proximal Point Algorithm." *SIAM Journal on Control and Optimization* 14 (5): 877–898. doi:10.1137/0314056. (Cited on pages 91, 93, 100, 107, 111).

Rockafellar, R. T., and R. J.-B. Wets (1998). *Variational analysis.* Vol. 317. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin. (Cited on page 155).

Rosenthal, J. S. (1995). "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo." *Journal of the American Statistical Association* 90 (430): 558–566. doi:10.1080/01621459.1995.10476548. (Cited on page 79).

Roueff, A., M. Gerin, P. Gratier, F. Levrier, J. Pety, M. Gaudel, J. R. Goicoechea, et al. (2020). "C18O, 13CO, and 12CO abundances and excitation temperatures in the Orion B molecular cloud: An analysis of the precision achievable when modeling spectral line within the Local Thermodynamic Equilibrium approximation." *Astronomy & Astrophysics (in press).* arXiv: 2005.08317. (Cited on page 25).

Rue, H. (2001). "Fast Sampling of Gaussian Markov Random Fields." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63 (2): 325–338. doi:10.1111/1467-9868.00288. (Cited on page 96).

Rue, H., and L. Held (2005). *Gaussian Markov Random Fields: Theory And Applications.* Chapman & Hall/CRC. (Cited on pages 96, 97, 98, 101, 104).

Ruszczynski, A. (1994). *A Partial Regularization Method for Saddle Point Seeking.* IIASA Working Paper. Available at http://pure.iiasa.ac.at/id/eprint/4189/, March. (Cited on page 112).

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems.* 2nd. Philadelphia, PA, USA: Society for Industrial / Applied Mathematics. (Cited on page 102).

Scheffé, H. (1947). "A useful convergence theorem for probability distributions." *The Annals of Mathematical Statistics* 18 (3): 434–438. doi:10.1214/aoms/1177730390. (Cited on page 30).

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016). "Bayes and Big Data: The Consensus Monte Carlo Algorithm." Available at https://research.google/pubs/pub41849/, *International Journal of Management Science and Engineering Management* 11:78–88. (Cited on page 34).

She, Y., and A. B. Owen (2011). "Outlier Detection Using Nonconvex Penalized Regression." *Journal of the American Statistical Association* 106 (494): 626–639. doi:10.1198/jasa.2011.tm10390. (Cited on page 36).

Shinozuka, M., and C.-M. Jan (1972). "Digital simulation of random processes and its applications." *Journal of Sound and Vibration* 25 (1): 111–128. doi:10.1016/0022-460X(72)90600-1. (Cited on page 97).

Sisson, S. A., Y. Fan, and M. A. Beaumont (2018a). "Overview of Approximate Bayesian Computation." In *Handbook of Approximate Bayesian Computation,* 1st ed., edited by Author, 3–54. Chapman / Hall/CRC Press. (Cited on pages 38, 39).

Sisson, S., Y. Fan, and M. Beaumont, eds. (2018b). *Handbook of Approximate Bayesian Computation.* 1st ed. Chapman / Hall/CRC Press. (Cited on pages 27, 36).

Starck, J.-L., F. Murtagh, and J. Fadili (2015). *Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis.* 2nd ed. Cambridge University Press. doi:10.1017/CBO9781316104514. (Cited on page 63).

Steidl, G., J. Weickert, T. Brox, P. Mrázek, and M. Welk (2004). "On the Equivalence of Soft Wavelet Shrinkage, Total Variation Diffusion, Total Variation Regularization, and SIDEs." *SIAM Journal on Numerical Analysis* 42 (2): 686–713. doi:10.1137/S0036142903422429. (Cited on page 73).

Steutel, F. W., and K. van Harn (2003). *Infinite Divisibility of Probability Distributions on the Real Line.* 1st ed. Pure and Applied Mathematics. CRC Press. (Cited on page 118).

Stone, C. J., and S. Wainger (1967). "One-sided error estimates in renewal theory." *Journal d'Analyse Mathématique* 20:325–352. doi:10.1007/BF02786679. (Cited on page 78).

Tanner, M. A., and W. H. Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398): 528–540. doi:10.1080/01621459.1987.10478458. (Cited on page 29).

Thomas, D. B., W. Luk, P. H. Leong, and J. D. Villasenor (2007). "Gaussian Random Number Generators." *ACM Computing Surveys* 39, no. 4 (November). doi:10.1145/1287620.1287622. (Cited on page 95).

Trench, W. (1964). "An Algorithm for the Inversion of Finite Toeplitz Matrices." *SIAM Journal on Applied Mathematics* 12, no. 3 (September): 512–522. doi:10.1137/0112045. (Cited on page 101).

Tripathi, G. (1999). "A matrix extension of the Cauchy-Schwarz inequality." *Economics Letters* 63 (1): 1–3. doi:10.1016/S0165-1765(99)00014-2. (Cited on page 175).

Tuominen, P., and R. L. Tweedie (1994). "Subgeometric Rates of Convergence of f-Ergodic Markov Chains." *Advances in Applied Probability* 26 (3): 775–798. doi:10.2307/1427820. (Cited on page 78).

Velayudhan, D., and S. Paul (2016). "Two-phase approach for recovering images corrupted by Gaussian-plus-impulse noise." In *International Conference on Inventive Computation Technologies,* 2:1–7. doi:10.1109/INVENTIVE.2016.7824875. (Cited on page 58).

Villani, C. (2008). *Optimal Transport: Old and New.* Springer Berlin Heidelberg. (Cited on pages 145, 151, 156, 159).

Vono, M., E. Bron, P. Chainais, F. L. Petit, S. Bardeau, S. Bourguignon, J. Chanussot, et al. (2019). "A fully Bayesian approach for inferring physical properties with credibility intervals from noisy astronomical data." In *Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing,* 1–5. doi:10.1109/WHISPERS.2019.8920859. (Cited on page 25).

Vono, M., N. Dobigeon, and P. Chainais (2018). "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler." In *IEEE International Workshop on Machine Learning for Signal Processing.* doi:10.1109/MLSP.2018.8516963. (Cited on pages 25, 36, 53, 130).

Vono, M., N. Dobigeon, and P. Chainais (2019a). "Split-and-augmented Gibbs sampler - Application to large-scale inference problems." *IEEE Transactions on Signal Processing* 67 (6): 1648–1661. doi:10.1109/TSP.2019.2894825. (Cited on pages 25, 32, 34, 39, 53, 104, 129).

Vono, M., N. Dobigeon, and P. Chainais (2019b). "Bayesian image restoration under Poisson noise and log-concave prior." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/ICASSP.2019.8683031. (Cited on pages 25, 54, 130).

Vono, M., N. Dobigeon, and P. Chainais (2019c). "Efficient sampling through variable splitting-inspired Bayesian hierarchical models." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/ICASSP.2019.8682982. (Cited on pages 25, 53, 130).

Vono, M., N. Dobigeon, and P. Chainais (2019d). "Modèles augmentés asymptotiquement exacts." In *Proc. of GRETSI*. (Cited on pages 25, 28, 129).

Vono, M., N. Dobigeon, and P. Chainais (2019e). "Un modèle augmenté asymptotiquement exact pour la restauration bayésienne d'images dégradées par un bruit de Poisson." In *Proc. of GRETSI*. (Cited on pages 25, 54, 130).

Vono, M., N. Dobigeon, and P. Chainais (2020a). "Asymptotically exact data augmentation: models, properties and algorithms." *Journal of Computational and Graphical Statistics (in press)*. doi:10.1080/10618600.2020.1826954. (Cited on pages 24, 28, 52, 129).

Vono, M., N. Dobigeon, and P. Chainais (2020b). "High-dimensional Gaussian sampling: A review and a unifying approach based on a stochastic proximal point algorithm." In 1st round of review, *SIAM Review*. arXiv: 2010.01510. (Cited on pages 24, 93, 100, 130).

Vono, M., D. Paulin, and A. Doucet (2019). "Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting." In 2nd round of review, *Journal of Machine Learning Research*. arXiv: 1905.11937. (Cited on pages 24, 26, 28, 77, 130).

Wang, C., and D. M. Blei (2018). "A General Method for Robust Bayesian Modeling." *Bayesian Analysis* 13 (4): 1163–1191. doi:10.1214/17-BA1090. (Cited on pages 35, 36).

Wang, X., and D. B. Dunson (2013). *Parallelizing MCMC via Weierstrass Sampler*. [online]. Technical report. Available at https://arxiv.org/abs/1312.4605/. technical report. (Cited on page 34).

Wang, Y., J. Yang, W. Yin, and Y. Zhang (2008). "A New Alternating Minimization Algorithm for Total Variation Image Reconstruction." *SIAM Journal on Imaging Sciences* 1 (3): 248–272. doi:10.1137/080724265. (Cited on pages 56, 62).

Welling, M., and Y. W. Teh (2011). "Bayesian Learning via Stochastic Gradient Langevin Dynamics." In *International Conference on International Conference on Machine Learning*, 681–688. Available at https://www.ics.uci.edu/\protect\unhbox\voidb@x\penalty\@M\{}welling/publications/papers/stoclangevin_v6.pdf. (Cited on page 21).

Wilhelm, S., and B. Manjunath (2010). "tmvtnorm: A Package for the Truncated Multivariate Normal Distribution." *The R Journal* 2 (June). doi:10.32614/RJ-2010-005. (Cited on page 98).

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. doi:10.1515/sagmb-2013-0010. (Cited on pages 31, 36).

Wood, A. T. A., and G. Chan (1994). "Simulation of Stationary Gaussian Processes in $[0, 1]^d$." *Journal of Computational and Graphical Statistics* 3 (4): 409–432. doi:10.1080/10618600.1994.10474655. (Cited on page 97).

Xu, X., and M. Ghosh (2015). "Bayesian Variable Selection and Estimation for Group Lasso." *Bayesian Analysis* 10 (4): 909–936. doi:10.1214/14-BA929. (Cited on page 81).

Zhang, X., M. Burger, and S. Osher (2011). "A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration." *Journal of Scientific Computing* 46, no. 1 (January): 20–46. doi:10.1007/s10915-010-9408-8. (Cited on page 113).