Gradient Scan Gibbs Sampler: an efficient algorithm for high-dimensional Gaussian distributions

Olivier Féron, François Orieux, Jean-François Giovannelli

Gradient Scan Gibbs Sampler: an efficient algorithm for high-dimensional Gaussian distributions

O. Féron*, F. Orieux and J.-F. Giovannelli

***Abstract*—This paper deals with Gibbs samplers that include high dimensional conditional Gaussian distributions. It proposes an efficient algorithm that avoids the high dimensional Gaussian sampling and relies on a random excursion along a small set of directions. The algorithm is proved to converge, *i.e.* the drawn samples are asymptotically distributed according to the target distribution. Our main motivation is in inverse problems related to general linear observation models and their solution in a hierarchical Bayesian framework implemented through sampling algorithms. It finds direct applications in semi-blind / unsupervised methods as well as in some non-Gaussian methods. The paper provides an illustration focused on the unsupervised estimation for super-resolution methods.**

I. Introduction

A. Context and problem statement

Gaussian distributions are common throughout signal and image processing, machine learning, statistics,... being convenient from both theoretical and numerical standpoints. Moreover, they are versatile enough to describe very diverse situations. Nevertheless, efficient sampling including these distributions is a cumbersome problem in high dimensions and this paper deals with this question.

Our main motivation is in inverse problems [1], [2] and the methodology resorts to a hierarchical Bayesian strategy, numerically implemented through Monte-Carlo Markov Chain algorithms and more specifically the Gibbs Sampler (GS). Indeed, consider the general linear direct model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$, where $\mathbf{y}$, $\mathbf{n}$ and $\mathbf{x}$ are the observation, the noise and the unknown image and $\mathbf{A}$ is a given linear operator. Consider, again, two independent prior distributions for $\mathbf{n}$ and $\mathbf{x}$ that are Gaussian conditionally to a vector $\boldsymbol{\theta}$, namely the hyper-parameter vector. The estimation of both $\mathbf{x}$ and $\boldsymbol{\theta}$ relies on the sampling of the joint posterior $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, and this is the core question of the paper. It commonly requires the handling of the high dimensional conditional posterior $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ that is Gaussian with given mean $\mathbf{m}$ and precision $\mathbf{Q}$.

The framework considered in this paper directly covers non-stationary and inhomogeneous Gaussian models for image and noise. The paper also has fallouts for non-Gaussian models based on conditionally Gaussian ones involving auxiliary / latent variables[1] (*e.g.*, location or scale mixtures of Gaussian) for edge preserving [3]–[5] and for sparse signals [6], [7]. It also includes other hierarchical models [8], [9] involving labels for inversion-segmentation. This framework also includes linear variant direct models and some non-linear direct models, based on conditional linear ones, *e.g.* bilinear or multilinear. In addition, it covers a majority of current inverse problems, *e.g.* unsupervised [5] and semi-blind [10], by including hyperparameters and acquisition parameters in the vector $\boldsymbol{\theta}$.

Large scale Gaussian distributions are also useful for Internet data processing, *e.g.* to model social networks and to develop recommender systems [11]. They are also widely used in epidemiology and disease mapping [12], [13] as they provide a simple way to include spatial correlations. The question is also in relation to spatial linear regression with (smooth) spatially varying parameters [14]. In these cases the question of efficient sampling including Gaussian distributions in high dimensions becomes crucial and it is all the more true in the "Big Data" context.

In the following we address the general problem of sampling from a joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$ where the conditional distribution $p(\mathbf{x}|\boldsymbol{\theta})$ is a high-dimensional Gaussian distribution.

B. Existing approaches

The difficulty is directly related to handling the high-dimensional precision $\mathbf{Q}$. The factorization (Cholesky, square root,...), diagonalization and inversion of $\mathbf{Q}$ could be used but they are generally unfeasible in high dimensions due to both computational cost and memory footprint. Nevertheless, such solutions are practicable in two famous cases.

- If $\mathbf{Q}$ is circulant or circulant-block-circulant an efficient strategy [15], [16] relies on its diagonalization computed by FFT. More generally, an efficient strategy exists if $\mathbf{Q}$ is diagonalizable by a fast transform, *e.g.* discrete cosine transform for Neumann boundary conditions [17], [18].
- When $\mathbf{Q}$ is sparse, a possible strategy [13], [19], [20] relies on a Cholesky decomposition and a linear system resolution. Another strategy is a GS [21] that simultaneously updates large blocks of variables.

O. Féron is with EDF Research & Developments, 92140 Clamart, France and with Univ. Paris Dauphine, FiME, 75116 Paris, France, olivier-2.feron@edf.fr. F. Orieux is with Univ. Paris-Sud 11, L2S, UMR 8506, 91190 Gif-sur-Yvette, France, orieux@l2s.centralesupelec.fr. J.-F. Giovannelli is with Univ. Bordeaux, IMS, UMR 5218, F-33400 Talence, France, Giova@IMS-Bordeaux.fr.

[1] It is based on the fact that for a couple of random variables (U, V), the conditional law for $U|V$ is Gaussian and the marginal law for U is non-Gaussian. A famous example is a Gaussian variable with precision under a Gamma distribution: the resulting marginal follow a Student distribution.

In order to address more general cases, solutions founded on iterative algorithms for objective optimization or linear system resolution have recently been proposed.

1) An efficient algorithm has been proposed by several authors [6], [17], [18], [22], [23] (previously used in applications [8], [10]). It is founded on a Perturbation-Optimization (PO) principle: adequate stochastic perturbation of a quadratic criterion and optimization of the perturbed criterion. However, in order to obtain a sample from the right distribution, an exact optimization is needed, but in practice an empirical truncation of the iterations is implemented, leading to an approximate sample. [24] introduces a Metropolis step in order to asymptotically retrieve an exact sample and then to ensure, in a global MCMC procedure, the convergence to the correct invariant distribution.
2) In [25], [26] the authors propose a Conjugate Direction Sampler (CDS) based on two crucial properties: (i) a Gaussian distribution admits Gaussian conditional distributions and (ii) a set of mutually conjugate directions w.r.t. $\mathbf{Q}$ is available. The key point of the algorithm is to sample along these mutually conjugate directions instead of optimizing as in the classical Conjugate Gradient optimization algorithm.

In the first case, the only constraint on $\mathbf{Q}$ is that a sample from $\mathcal{N}(0, \mathbf{Q})$ must be accessible, which is often the case in inverse problem applications. In the second case, $\mathbf{Q}$ must have only distinct eigenvalues to make the CDS give an exact sample. Otherwise it leads to an approximate sample as described in [26].

The proposed algorithm uses the same approach as the CDS and extends the efficiency to, theoretically, any matrix $\mathbf{Q}$.

C. Contribution

The existing methods described above and the proposed one are both founded on a Gibbs sampler. However, the existing ones attempt to sample the high dimensional Gaussian component $\mathbf{x} \in \mathbb{R}^N$ whereas the proposed method does not. Our main contribution is to avoid the high dimensional sampling and only requires small dimensional sampling. More precisely, given a subspace $D \subset \mathbb{R}^N$, the objective is to sample the sub-component of $\mathbf{x}$ according to the subspace D. It must be sampled under the appropriate conditional distribution $\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D}, \boldsymbol{\theta})$, with the decomposition $\mathbf{x} = (\mathbf{x}_D, \mathbf{x}_{\setminus D})$. The algorithm takes advantage of the ease of calculating the conditional pdf of a multivariate Gaussian distribution, when D is appropriately built, as explained in section II. These ideas are strongly related to other existing works.

- If the subset D is composed of only one direction in the canonical coordinates, the algorithm amounts to a pixel-by-pixel GS [3].
- The marginal chain $\mathbf{x}^{(t)}$ can also be viewed as the one produced by a specific random scan sampler [27]–[29]. The random scans are related to the random choice of D, depending on the current value $\boldsymbol{\theta}^{(t)}$.
- Other algorithms based on optimization principles [26], [30] aim at producing a complete optimization. On the the other hand, in essence, the proposed approach only requires a few steps of the optimization process.
- A similar idea is at work in Hamiltonian (or Langevin) Monte Carlo [31]–[34] (see also [35]): the proposed distribution takes advantage of an ascent direction of the target to increase the acceptation probability. Here, the exact distribution is sampled, so the proposal is always accepted.

However, to our knowledge, the proposed algorithm does not directly join the class of existing strategies. One contribution of this paper is to give sufficient assumptions for convergence, *i.e.* the samples are asymptotically distributed according to the joint pdf $p(\mathbf{x}, \boldsymbol{\theta})$.

D. Outline

Subsequently, Section II presents the proposed algorithm and section III gives an illustration through an academic problem in super-resolution. Section IV presents conclusions and perspectives.

II. Gradient Scan Gibbs Sampler

In this section we describe the proposed algorithm: a GS with a high dimensional conditional Gaussian distribution. The objective is to generate samples from a joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$, where $\mathbf{x} \in \mathbb{R}^N$ is highly dimensional and $p(\mathbf{x}|\boldsymbol{\theta})$ is a Gaussian distribution $\mathcal{N}(\mathbf{m}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}}^{-1})$:

$$p(\mathbf{x}|\boldsymbol{\theta}) = (2\pi)^{-N/2}(\det \mathbf{Q}_{\boldsymbol{\theta}})^{1/2} \exp -J_{\boldsymbol{\theta}}(\mathbf{x}) \tag{1}$$

with the potential $J_{\boldsymbol{\theta}}$ defined as:

$$J_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{m}_{\boldsymbol{\theta}})^{\mathrm{t}}\mathbf{Q}_{\boldsymbol{\theta}}(\mathbf{x} - \mathbf{m}_{\boldsymbol{\theta}}). \tag{2}$$

All the other variables of the problem are grouped into $\boldsymbol{\theta} \in \Theta$ and we assume that the sampling from $p(\boldsymbol{\theta}|\mathbf{x})$ is tractable (directly or with several steps of the GS, including Metropolis-Hastings steps).

A. Preliminary results

This section presents classical definitions and results, mostly based on [25], needed to provide convergence proof and links between matrix factorization and optimization / sampling procedures.

Definition 1. *Consider $\mathbf{Q}$ a $N \times N$ symmetric definite positive matrix. A set $\{\mathbf{d}_n, n = 1, \ldots, N\}$ of non-zero vectors in $\mathbb{R}^N$ such that: $\mathbf{d}_n^{\mathrm{t}}\,\mathbf{Q}\,\mathbf{d}_m = 0$ for $n, m = 1, \ldots, N,\ n \neq m$ is said mutually conjugate w.r.t. $\mathbf{Q}$.* △

A mutually conjugate set $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ w.r.t. $\mathbf{Q}$ is a basis of $\mathbb{R}^N$, then, for all $\mathbf{x} \in \mathbb{R}^N$:

$$\mathbf{x} = \sum_{n=1}^{N} \alpha_n \mathbf{d}_n \quad \text{with} \quad \alpha_n = \frac{\mathbf{d}_n^{\mathrm{t}}\mathbf{Q}\mathbf{x}}{\mathbf{d}_n^{\mathrm{t}}\mathbf{Q}\mathbf{d}_n}\,.$$

So, if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{Q}^{-1})$ is a Gaussian random vector with mean $\mathbf{m}$ and precision $\mathbf{Q}$, then the α_n are also Gaussian:

$$\alpha_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n^{\mathrm{t}}\mathbf{Q}\mathbf{m}}{\mathbf{d}_n^{\mathrm{t}}\mathbf{Q}\mathbf{d}_n}\ ;\ \frac{1}{\mathbf{d}_n^{\mathrm{t}}\mathbf{Q}\mathbf{d}_n}\right) \tag{3}$$

and reciprocally if the α_n are distributed under (3) then $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{Q}^{-1})$.

In particular, let $\mathbf{x}^0 \in \mathbb{R}^N$ be a "current" point and $\mathbf{d}_1 \in \mathbb{R}^N$ a given "direction". One can find $\mathbf{d}_2, \ldots, \mathbf{d}_N$ such that $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ is mutually conjugate w.r.t. $\mathbf{Q}$ and $\mathbf{x}^0$ writes:

$$\mathbf{x}^0 = \sum_{n=1}^{N} \alpha_n^0 \, \mathbf{d}_n.$$

Consider now the N_D-dimensional subset

$$D(\mathbf{x}^0) = \left\{ \sum_{n=1}^{N} \alpha_n \mathbf{d}_n, \alpha_n \in \mathbb{R},\ n \leq N_D, \alpha_n = \alpha_n^0,\ n > N_D \right\}$$
$$= \left\{ \mathbf{x}^0 + \sum_{n=1}^{N_D} (\alpha_n - \alpha_n^0)\mathbf{d}_n,\ (\alpha_1, \ldots, \alpha_{N_D}) \in \mathbb{R}^{N_D} \right\}$$

We are interested in the conditional pdf $p(\mathbf{x}|\mathbf{x} \in D(\mathbf{x}^0))$. The following result and its proof can be found in [25].

Proposition 1. *A sample $\widetilde{\mathbf{x}}$ according to $p(\mathbf{x}|\mathbf{x} \in D(\mathbf{x}^0))$ can be obtained by:*

1) sample independently the set $(\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_{N_D})$ with:

$$\widetilde{\alpha}_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q}(\mathbf{x}^0 - \mathbf{m})}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n} \ ; \ \frac{1}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n} \right), n = 1, \ldots, N_D$$

2) compute $\widetilde{\mathbf{x}} = \mathbf{x}^0 - \sum_{n=1}^{N_D} \widetilde{\alpha}_n \, \mathbf{d}_n$

B. Gradient Scan Gibbs Sampler (GSGS)

In the following we propose a GS in order to sample the joint probability $p(\mathbf{x}, \boldsymbol{\theta})$. The principle is to sample, at each iteration of the GS, only N_D directions of $\mathbf{x}$ instead of sampling the whole high dimensional variable. The chosen first direction of the set D will be the gradient of the potential of $p(\mathbf{x}|\boldsymbol{\theta})$, with a stochastic perturbation to ensure, in the general case, the convergence of the resulting Markov chain. The following directions are chosen so as to get a mutually conjugate subset with respect to the precision of $p(\mathbf{x}|\boldsymbol{\theta})$.

We call our proposed algorithm the Gradient Scan Gibbs Sampler (GSGS) which is described by Algorithm 1. In this algorithm the chosen first sampling direction $\mathbf{d}_1$ is given by the gradient of the potential of $p(\mathbf{x}|\boldsymbol{\theta})$, with an additional random perturbation $\tilde{\boldsymbol{\varepsilon}}$ that follows a probability density $p(\boldsymbol{\varepsilon})$. In fact, we expect the gradient to be a good direction towards regions of high probabilities. Also, the gradient is easily computable and so gives an easy rule to sample from any current point $\mathbf{x}$. Moreover, the other conjugate directions are iteratively computable as described in the Conjugate Direction Sampling (CDS) algorithm [25] used to get an approximated sample from a Gaussian distribution. In fact, the GSGS is embedding steps of the CDS in a global GS.

The objective is now to study the convergence properties of the GSGS. We begin with two classical results.

- If the Markov chain is aperiodic, $\phi-$irreducible for some nonzero measure ϕ[2], and has an invariant probability π, then it converges to π from π-almost every starting point (cf. Theorem 4.4 of [36]).
- Moreover, if the Markov chain is Harris recurrent, then it converges to π from all starting point [36], [37].

[2]In all the paper we will consider ϕ as the Lebesgue measure and we will omit it for simplicity.

Algorithm 1 : Gradient scan Gibbs sampler (GSGS).

Define an initial point $\mathbf{x}^{(0)}$, a number N_D and a stopping criterion. Iterate .

1: sample $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta}|\mathbf{x}^{(t-1)})$

2: set $\mathbf{Q}_t = \mathbf{Q}_{\boldsymbol{\theta}^{(t)}}$ and $\mathbf{m}_t = \mathbf{m}_{\boldsymbol{\theta}^{(t)}}$, and compute the gradient $\mathbf{g} = \nabla J_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}) = \mathbf{Q}_t(\mathbf{x}^{(t-1)} - \mathbf{m}_t)$

3: sample a perturbation $\tilde{\boldsymbol{\varepsilon}} \sim p(\boldsymbol{\varepsilon})$

4: compute a set of N_D mutually conjugate directions $(\mathbf{d}_1, \ldots, \mathbf{d}_{N_D})$ w.r.t. $\mathbf{Q}_t$ such that

$$\mathbf{d}_1 = \mathbf{g} + \tilde{\boldsymbol{\varepsilon}}$$

5: sample independently the set $(\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_{N_D})$ with:

$$\widetilde{\alpha}_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n^{\mathrm{t}} \mathbf{g}}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q}_t \mathbf{d}_n} \ ; \ \frac{1}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q}_t \mathbf{d}_n} \right), \ n \leq N_D$$

6: compute $\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \sum_{n=1}^{N_D} \widetilde{\alpha}_n \, \mathbf{d}_n$

7: $t \leftarrow t + 1$.

until the stopping criterion is reached.

The Harris recurrence of GS, or more generally Metropolis-within-Gibbs samplers is well studied in [37]. In particular, the Theorem 12 and Corollary 13 of [37] ensures that if the Markov chain produced by the GSGS is irreducible then it is Harris recurrent. Consequently, in the following we focus on showing that the Markov chain is aperiodic, irreducible and with stationary distribution $p(\mathbf{x}, \boldsymbol{\theta})$.

It is trivial to see that the Markov chain $(\mathbf{x}^{(t)}, \boldsymbol{\theta}^{(t)})_{t \geq 0}$, produced by the GSGS, is aperiodic since for any non-negligible subset $A \in \mathbb{R}^N$ including $\mathbf{x}^{(t-1)}$, $\mathbb{P}(\mathbf{x}^{(t)} \in A) > 0$. The existence of an invariant probability and the irreducibility can be shown by thinking of a random scan GS for the marginal component $(\mathbf{x}^{(t)})_{t \geq 0}$.

Proposition 2. *The Markov chain produced by Algorithm 1 admits $p(\mathbf{x}, \boldsymbol{\theta})$ as an invariant distribution, even without perturbations of the gradient direction (i.e. $\tilde{\boldsymbol{\varepsilon}} = 0$).*
Moreover, if the density $p(\boldsymbol{\varepsilon})$ is supported on $\mathbb{R}^N$, the Markov chain produced by Algorithm 1 is irreducible, and therefore its law converges to $p(\mathbf{x}, \boldsymbol{\theta})$.

Proof. see appendix A. □

Proposition 2 then shows that the joint probability $p(\mathbf{x}, \boldsymbol{\theta})$ remains an invariant distribution in the limit case where the first direction $\mathbf{d}_1$ is exactly the gradient of $p(\mathbf{x}|\boldsymbol{\theta})$, without random perturbation. However the perturbation is needed to ensure the irreducibility (and then the convergence) of the chain.

If the gradient is not perturbed, the mutually conjugate set D is then given by a deterministic function of $\boldsymbol{\theta}^{(t)}$ and

$\mathbf{x}^{(t-1)}$. In this case, we need more assumptions to ensure the Markov chain to be irreducible. For example, we can have the following result.

Proposition 3. *Suppose the following conditions are satisfied:*

H-1 The function $\boldsymbol{\theta} \mapsto \mathbf{Q}_{\boldsymbol{\theta}}$ *is continuous*

H-2 $\forall(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^N \times \Theta$ *and* $\forall r > 0$, $\mathbb{P}(\mathcal{B}(\boldsymbol{\theta}, r)|\mathbf{x}) > 0$, *with* $\mathcal{B}(\boldsymbol{\theta}, r)$ *the ball in* Θ, *centered in* $\boldsymbol{\theta}$, *of radius* r.

H-3 $\forall \mathbf{x} \in \mathbb{R}^N$, $\exists \boldsymbol{\theta} \in \Theta$ *such as:*

H-3.1 $\mathbf{Q}_{\boldsymbol{\theta}}$ *has* N *distinct eigenvalues,*

H-3.2 $\mathbf{x} - \mathbf{m}_{\boldsymbol{\theta}}$ *is not orthogonal to any eigenvector of* $\mathbf{Q}_{\boldsymbol{\theta}}$,

Then the Markov chain produced by Algorithm 1 without the perturbation step 3 ($\tilde{\boldsymbol{\varepsilon}} = 0$*) is irreducible.*

Proof. see appendix B □

The conditions described in Proposition 3 are very restrictive and, in particular, condition H-3.1 is difficult, if not impossible, to prove in practice. This condition ensures that every non-negligible subset of $\mathbb{R}^N$ can be reached with a non-zero probability. It can be interpreted in the framework of Krylov spaces as in [26]. For example, if there is t such as the Krylov space

$$\mathcal{K}^N(\mathbf{Q}_{\boldsymbol{\theta}^{(t)}}, \mathbf{x}^{(t)}) := \text{span}\left(\mathbf{x}^{(t)}, \mathbf{Q}_{\boldsymbol{\theta}^{(t)}}\mathbf{x}^{(t)}, \ldots, \mathbf{Q}^N_{\boldsymbol{\theta}^{(t)}}\mathbf{x}^{(t)}\right)$$

is of rank N then the Markov chain is irreducible. This condition can be weakened in our case because the Gaussian parameters $\mathbf{m}_{\boldsymbol{\theta}^{(t)}}$ and $\mathbf{Q}_{\boldsymbol{\theta}^{(t)}}$ are changing since $\boldsymbol{\theta}$ is changing at each iteration of the GS. Therefore a sufficient condition to ensure the irreducibility of the chain can be expressed as follows:

Proposition 4. *If there is* $T > N$ *such as the union of Krylov spaces*

$$\cup_{t=1}^{T} \mathcal{K}^N(\mathbf{Q}^{(t)}_{\boldsymbol{\theta}}, \mathbf{x}^{(t)}) \cup \mathcal{K}^N(\mathbf{Q}^{(t)}_{\boldsymbol{\theta}}, \mathbf{m}^{(t)}_{\boldsymbol{\theta}})$$

is of rank N *then the Markov chain built by the GSGS without perturbation of the gradient is irreducible.*

Proof. The condition implies that for any non-negligible subset $A \subset \mathbb{R}^N$, $\mathbb{P}\left(\mathbf{x}^{(T)} \in A|\mathbf{x}^{(0)}\right) > 0$, which ensures the irreducibility. □

The issue of determining general conditions, as in Proposition 3, is an open problem at this time. The fact that the condition described in Proposition 4 is satisfied, highly depends on the model's characteristics. That is why the GSGS (with the random perturbation step 3) is the one that ensures, in all cases, the convergence of the Markov chain to the joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$.

The above results do not allow us to get any convergence rate of the Markov chain. The latter is, in fact, very important to ensure in practice the efficiency of the estimators produced by simulations in finite time. In particular, the geometric ergodicity [38] is a very well known property that gives a Central Limit Theorem and ensures the Markov chain to quickly converge and give estimations of standard errors. However the Algorithm 1 aims to be general while the precise study of geometric convergence (especially to quantify the convergence rate) would need to specify the distributions on the parameters $\boldsymbol{\theta}$ and on the perturbation $\boldsymbol{\varepsilon}$. At this time, only weak assumptions are considered on these probabilities and the next section discusses the different choices of $p(\boldsymbol{\varepsilon})$ from a feasibility point of view.

C. Choice of $p(\boldsymbol{\varepsilon})$

As previously specified, the only condition to ensure the convergence of the GSGS in the general case, is to choose a distribution $p(\boldsymbol{\varepsilon})$ supported in $\mathbb{R}^N$. In practice we also expect a sample from $p(\boldsymbol{\varepsilon})$ to be easily accessible. A natural choice is the Gaussian iid distribution $\mathcal{N}(0, \mathbf{I}_N)$, $\mathbf{I}_N$ being the $N \times N$ identity matrix. This was already studied in [39] in the case of only sampling from a Gaussian distribution $p(\mathbf{x})$ and where results are shown in small dimensions.

Our empirical studies in high dimension (one example is shown in section III) incited us to choose the Gaussian distribution $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$, when it is possible. The sampling from this distribution may actually be easily computable, provided that $\mathbf{Q}_{\boldsymbol{\theta}}$ has, for example, the specific factorization form described in [30]:

$$\mathbf{Q}_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \mathbf{M}_k^{\mathrm{t}} \mathbf{R}_k^{-1} \mathbf{M}_k$$

In this case, the sampling from $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$ is easily computable by using the Perturbation Optimization (PO) algorithm [30]. The latter consists in (i) randomly modifying the potential $J_{\boldsymbol{\theta}}(\mathbf{x})$ to get a perturbed potential $\tilde{J}_{\boldsymbol{\theta}}$ and (ii) optimizing $\tilde{J}_{\boldsymbol{\theta}}$. The first step of this optimization procedure consists in computing the gradient $\nabla \tilde{J}_{\boldsymbol{\theta}}$ and it is trivial to show that it can be decomposed: $\nabla \tilde{J}_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla J_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$. Therefore, the perturbed gradient $\mathbf{d}_1$ of the GSGS, with a random perturbation $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$, can be obtained by using the PO algorithm truncated to one step of the optimization procedure.

Although, at this time, this choice is empirical we may have some intuition to recommend, when it is possible, the distribution $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$. The first direction $\mathbf{d}_1$ is related to the gradient of $J_{\boldsymbol{\theta}}$, in accordance with the objective to get a direction towards regions of high probability. This gradient is mostly driven by the highest eigenvalues of $\mathbf{Q}_{\boldsymbol{\theta}}$. The perturbation $\boldsymbol{\varepsilon}$ is only needed to ensure the GSGS convergence, but the objective is to keep a direction towards high probability regions. The sampling from $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$ seems to be a good compromise: it gives values of $\boldsymbol{\varepsilon}$ mostly driven by the highest eigenvalues of $\mathbf{Q}_{\boldsymbol{\theta}}$ and then the resulting direction $\mathbf{d}_1$ still continues to encourage the exploration space of high probability.

We may also notice that some relaxations of the GSGS are possible, following classical arguments of a random scan GS. For example, it is not necessary to sample the perturbation from $p(\boldsymbol{\varepsilon})$ at each iteration, it is sufficient to do this an infinite number of times to ensure the chain to be irreducible.[3] As we will see in section III, a low frequency sampling of $\boldsymbol{\varepsilon}$ can improve the algorithm's efficiency.

[3]From any point $\left(\mathbf{x}^{(t)}, \boldsymbol{\theta}^{(t)}\right)$, let $s > t$ be the closest next time where $\boldsymbol{\varepsilon}$ is sampled, then for any non-negligible subset $A \in \mathbb{R}^N \times \boldsymbol{\Theta}$, we have $P(\mathbf{x}^{(t)}, A) > 0$.

III. Unsupervised super resolution as a large scale problem

A. Problem statement

The paper details an application of the proposed GSGS to a super-resolution problem (identical to the one presented in [30], [40]): several blurred, noisy and down-sampled (low resolution) observations of a scene are available to retrieve the original (high resolution) scene [41], [42].

The usual direct model reads: $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} = \mathbf{S}\mathbf{H}\mathbf{x} + \mathbf{n}$. In this equation, $\mathbf{y} \in \mathbb{R}^M$ collects the pixels of the low resolution images (five 128×128 images, *i.e.* $M = 81920$) and $\mathbf{x} \in \mathbb{R}^N$ collects the pixels of the original image (one 256×256 image, *i.e.* $N = 65536$). The noise $\mathbf{n} \in \mathbb{R}^M$ accounts for measurement and modeling errors. $\mathbf{H}$ is a $N \times N$ circulant-block-circulant convolution matrix accounting for the optical and the sensor parts of the observation system. Here it is a square window of 5-pixel-width. $\mathbf{S}$ is a $M \times N$ matrix modeling motion (here translation) and decimation: it is a down-sampling binary matrix indicating which pixel of the blurred image is observed.

The noise is chosen to be $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \gamma_{\mathbf{n}}^{-1}\mathbf{I})$. Regarding the object, the chosen prior accounts for smoothness: $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \gamma_{\mathbf{x}}^{-1}\mathbf{D}^{\mathrm{t}}\mathbf{D})$ where $\mathbf{D}$ is the $N \times N$ circulant convolution matrix of the Laplacian filter. The hyperparameters $\gamma_{\mathbf{n}}$ and $\gamma_{\mathbf{x}}$ are unknown and the assigned priors are conjugate : Gamma distributions $\gamma_{\mathbf{n}} \sim \mathcal{G}(\alpha_{\mathbf{n}}; \beta_{\mathbf{n}})$ and $\gamma_{\mathbf{x}} \sim \mathcal{G}(\alpha_{\mathbf{x}}; \beta_{\mathbf{x}})$. They are weakly informative for large variances and uninformative Jeffreys' prior when the $(\alpha_{\mathbf{x}}, \beta_{\mathbf{x}})$ tends to $(0, 0)$. As a consequence, the full posterior pdf writes

$$\begin{aligned} p(\mathbf{x}, \gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x}, \gamma_{\mathbf{n}}) p(\mathbf{x}|\gamma_{\mathbf{x}}) p(\gamma_{\mathbf{x}}) p(\gamma_{\mathbf{n}}) \\ &\propto \gamma_{\mathbf{n}}^{\alpha_{\mathbf{n}}+N/2-1} \gamma_{\mathbf{x}}^{\alpha_{\mathbf{x}}+(M-1)/2-1} \\ &\quad \exp\left[-\gamma_{\mathbf{n}}\|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}\|^2/2\right] \exp\left[-\beta_{\mathbf{n}}\gamma_{\mathbf{n}}\right] \\ &\quad \exp\left[-\gamma_{\mathbf{x}}\|\mathbf{D}\mathbf{x}\|^2/2\right] \exp\left[-\beta_{\mathbf{x}}\gamma_{\mathbf{x}}\right]. \end{aligned} \tag{4}$$

The conditional law of the image writes

$$p(\mathbf{x}|\mathbf{y}, \gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}) \propto \exp\left[-\frac{\gamma_{\mathbf{n}}}{2}\|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}\|^2 - \frac{\gamma_{\mathbf{x}}}{2}\|\mathbf{D}\mathbf{x}\|^2\right].$$

Accordingly the negative logarithm gives the criterion

$$J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}(\mathbf{x}) = \frac{\gamma_{\mathbf{n}}}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \frac{\gamma_{\mathbf{x}}}{2}\|\mathbf{D}\mathbf{x}\|^2$$

and the gradient

$$\begin{aligned} \nabla J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}(\mathbf{x}) &= \gamma_{\mathbf{n}}\mathbf{A}^{\mathrm{t}}(\mathbf{A}\mathbf{x} - \mathbf{y}) + \gamma_{\mathbf{x}}\mathbf{D}^{\mathrm{t}}\mathbf{D}\mathbf{x} \\ &= \mathbf{Q}(\mathbf{x} - \mathbf{m}) \end{aligned}$$

with $\mathbf{m} = \mathbf{Q}_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}^{-1}\gamma_{\mathbf{n}}\mathbf{A}^{\mathrm{t}}\mathbf{y}$, and the Hessian

$$\mathbf{Q}_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}} = \nabla^2 J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}(\mathbf{x}) = \gamma_{\mathbf{n}}\mathbf{A}^t\mathbf{A} + \gamma_{\mathbf{x}}\mathbf{D}^{\mathrm{t}}\mathbf{D}$$

B. Gibbs sampler

The posterior pdf is explored by the proposed GS in Algorithm 2, based on the GSGS, that iteratively updates $\gamma_{\mathbf{n}}$, $\gamma_{\mathbf{x}}$ and a sub-component of $\mathbf{x}$. Regarding the hyperparameters, the conditional pdf are Gamma and their parameters are easy to compute.

Algorithm 2 : GSGS for super-resolution.

Set $t = 1$, define an initial point $\mathbf{x}^{(0)}$, and repeat

1: Sample $\gamma_{\mathbf{n}}^{(t)} \sim p\left(\gamma_{\mathbf{n}}|\mathbf{y}, \mathbf{x}^{(t-1)}\right)$ as

$$\mathcal{G}\left(\frac{N}{2}; \frac{2}{\|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}^{(t-1)}\|^2}\right).$$

and $\gamma_{\mathbf{x}}^{(t)} \sim p\left(\gamma_{\mathbf{x}}|\mathbf{y}, \mathbf{x}^{(t-1)}\right)$ as

$$\mathcal{G}\left(\frac{M-1}{2}; \frac{2}{\|\mathbf{D}\mathbf{x}^{(t-1)}\|^2}\right).$$

2: Set $\mathbf{Q}_t = \mathbf{Q}_{\gamma_{\mathbf{x}}^{(t)}, \gamma_{\mathbf{n}}^{(t)}}$ and compute the gradient

$$\mathbf{g}^{(t)} = \nabla J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}\left(\mathbf{x}^{(t-1)}\right) = \mathbf{Q}_t(\mathbf{x}^{(t-1)} - \mathbf{m})$$

3: Sample a perturbation $\boldsymbol{\varepsilon}^{(t)} \sim \mathcal{N}(0, \mathbf{Q}_t)$

4: Compute a set of N_D mutually conjugate directions $\{\mathbf{d}_1, \ldots, \mathbf{d}_{N_D}\}$ with the first being $\mathbf{d}_1 = \mathbf{g}^{(t)} + \boldsymbol{\varepsilon}^{(t)}$.

5: Sample independently the set $(\widetilde{\alpha}_n)_{n=1,\ldots,N_D}$ with:

$$\widetilde{\alpha}_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n\mathbf{g}^{(t)}}{\mathbf{d}_n\mathbf{Q}_t\mathbf{d}_n}; \frac{1}{\mathbf{d}_n\mathbf{Q}_t\mathbf{d}_n}\right)$$

6: Compute $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)} - \sum_{n=1}^{N_D} \widetilde{\alpha}_n \mathbf{d}_n$.

7: $t \leftarrow t + 1$.

until the stopping criterion is reached.

The set of mutually conjugate directions w.r.t. $\mathbf{Q}_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}$, at step 4 of Algorithm 2, is computed by the Gram-Schmidt process applied to gradient, as usually found in conjugated gradient optimization algorithm. The procedure is similar to the algorithm described in [26]. Finally the estimator is the posterior mean computed as the empirical mean of the samples.

Despite the convergence proof with almost any law for the perturbation $\boldsymbol{\varepsilon}$ (provided that the density $p(\boldsymbol{\varepsilon})$ is supported in $\mathbb{R}^N$), some tuning is necessary to practically obtain a good space's exploration. In practice, Step 3 has a major influence and, as already discussed in section II-C, we observe that a working perturbation corresponds to those of the PO algorithm [30]

$$\boldsymbol{\varepsilon}^{(t)} = {\gamma_{\mathbf{n}}^{(t)}}^{-1/2}\mathbf{A}^{\mathrm{t}}\boldsymbol{\varepsilon}_{\mathbf{n}} + {\gamma_{\mathbf{x}}^{(t)}}^{-1/2}\mathbf{D}^{\mathrm{t}}\boldsymbol{\varepsilon}_{\mathbf{x}}$$

where $\boldsymbol{\varepsilon}_{\times}$ are two Gaussian normalized random vectors, leading to a Gaussian perturbation $\boldsymbol{\varepsilon}^{(t)}$ of covariance $\mathbf{Q}_t$. However, the proposed algorithm has numerous advantages over the PO algorithm. First the proposed algorithm has a convergence proof because it does not suffer from truncation, even in the extreme case with $N_D = 1$. Second the perturbation has the sole constraint of having $\mathbb{R}^N$ as support. Moreover a perturbation is not required at each iteration.

C. Numerical results

The posterior law (4) has been explored with the following four algorithms or settings.

- The adaptive RJ-PO algorithm [40], directly tuned with the acceptance probability, here chosen to be 0.9. This acceptance probability leads to an average number of

	PO	RJ-PO	GSGS(150)	GSGS(20)	GSGS(2)
$\widehat{\gamma_{\mathbf{n}}}$	0.9725	0.9718	0.9694	0.9694	0.7078
$\widehat{\sigma_{\gamma_{\mathbf{n}}}}$	0.0061	0.0063	0.0061	0.0063	0.0062
$\widehat{\gamma_{\mathbf{x}}}$	1.05 e-03	1.07e-03	1.06e-03	1.29e-03	9.62e-03
$\widehat{\sigma_{\gamma_{\mathbf{x}}}}$	1.5e-05	3.7e-05	1.7e-05	2.4e-05	6.2e-03
loop [s.]	3.4	2.4	2.4	0.5	0.1
total [s.]	515	362	353	72	9

Table I: Hyperparameter estimates and estimation variances for $\gamma_{\mathbf{n}} = 1$.

around 150 iterations of the conjugate gradient algorithm to compute the proposal, and with 6% of rejected samples.

- The PO algorithm [30] with a number of 150 iterations for the optimization.
- Algorithm 2 with $N_D = 150$. The idea is to build an algorithm close to RJ-PO's computing time.
- Algorithm 2 with $N_D = 20$. The idea is to show that our algorithm offers the possibility to reduce the number of iterations while still offering a good exploration and with guaranteed convergence.
- Algorithm 2 with $N_D = 2$. The idea is to show a very fast algorithm that offers a partially correct exploration. This case is particular in the sense that the perturbation is done only once for the whole algorithm.

The posterior mean (PM) estimations of the high-resolution image are given in Fig. 1 as well as the posterior standard deviation (PSD). From these results we can say that all algorithms provide similar quality for the image estimation. The same statement can be made for the standard deviation. However the posterior standard deviation with $N_D = 2$ seems incorrect. A possible interpretation is that the perturbation vector $\boldsymbol{\varepsilon}$ is simulated only once during the whole algorithm. Thus, the space is surely not sufficiently explored and the covariance estimation is severely biased. Indeed, since $\boldsymbol{\varepsilon}_{\times}$ are drawn only once, the stochastic explorations are limited to the conjugate direction plus the two directions $\boldsymbol{\varepsilon}_{\mathbf{x}}$ and $\boldsymbol{\varepsilon}_{\mathbf{n}}$. However the mean estimation does not seem to be affected and this algorithm is able to provide very quickly a good estimation of the image and hyperparameter values. We must notice that in our test with $N_D = 10$ the chain converged to a close, but wrong distribution, giving good results in the image but an slightly underestimation of $\gamma_{\mathbf{n}}$.

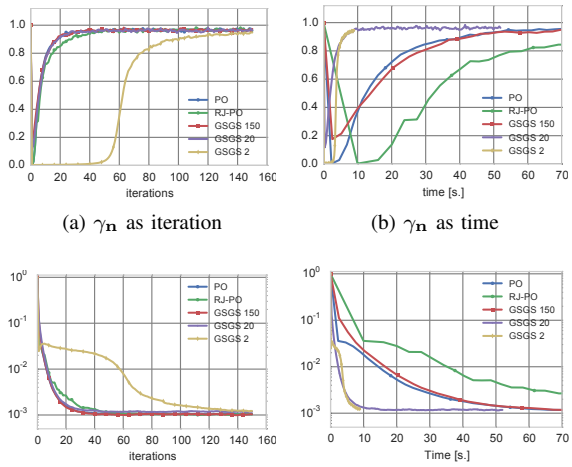
The chains of the hyperparameters are illustrated in Fig. 2. Figs. 2a and 2c represent the samples as a function of the iterations. We observe that, except for $N_D = 2$, all the chains have the same behavior with the same convergence period. The $N_D = 2$ has slower (in terms of the number of iterations) convergence but reaches the same stationary distribution.

Figs. 2b and 2d represent the samples as a function of time (in seconds). The chain behavior of algorithms PO, RJ-PO and GSGS(150) is very similar. This result is obvious since these algorithms compute almost the same number of gradients per iteration. That said, we see that for $N_D = 20$ and $N_D = 2$, the impact on the convergence time is significant. Table I shows some quantitative results. In particular the case $N_D = 20$ is five times faster than RJ-PO.

	PO	RJ-PO	GSGS(20)	GSGS(2)
$\widehat{\gamma_{\mathbf{n}}}$	9.9e-03	9.9e-03	9.9e-03	9.9e-03
$\widehat{\sigma_{\gamma_{\mathbf{n}}}}$	6.0e-05	6.05e-05	4.8e-05	5.5e-05
$\widehat{\gamma_{\mathbf{x}}}$	1.86e-03	1.84e-03	4.86e-03	2.29e-03
$\widehat{\sigma_{\gamma_{\mathbf{x}}}}$	3.2e-04	3.2e-04	7.2e-04	3.4e-05

Table II: Hyperparameter estimates and estimation variances for $\gamma_{\mathbf{n}} = 1e-02$.

In addition, Table II shows the estimated values of the hyperparameters with a higher noise level. Again the results are close with a good estimation of $\gamma_{\mathbf{n}}$.



(a) $\gamma_{\mathbf{n}}$ as iteration (b) $\gamma_{\mathbf{n}}$ as time

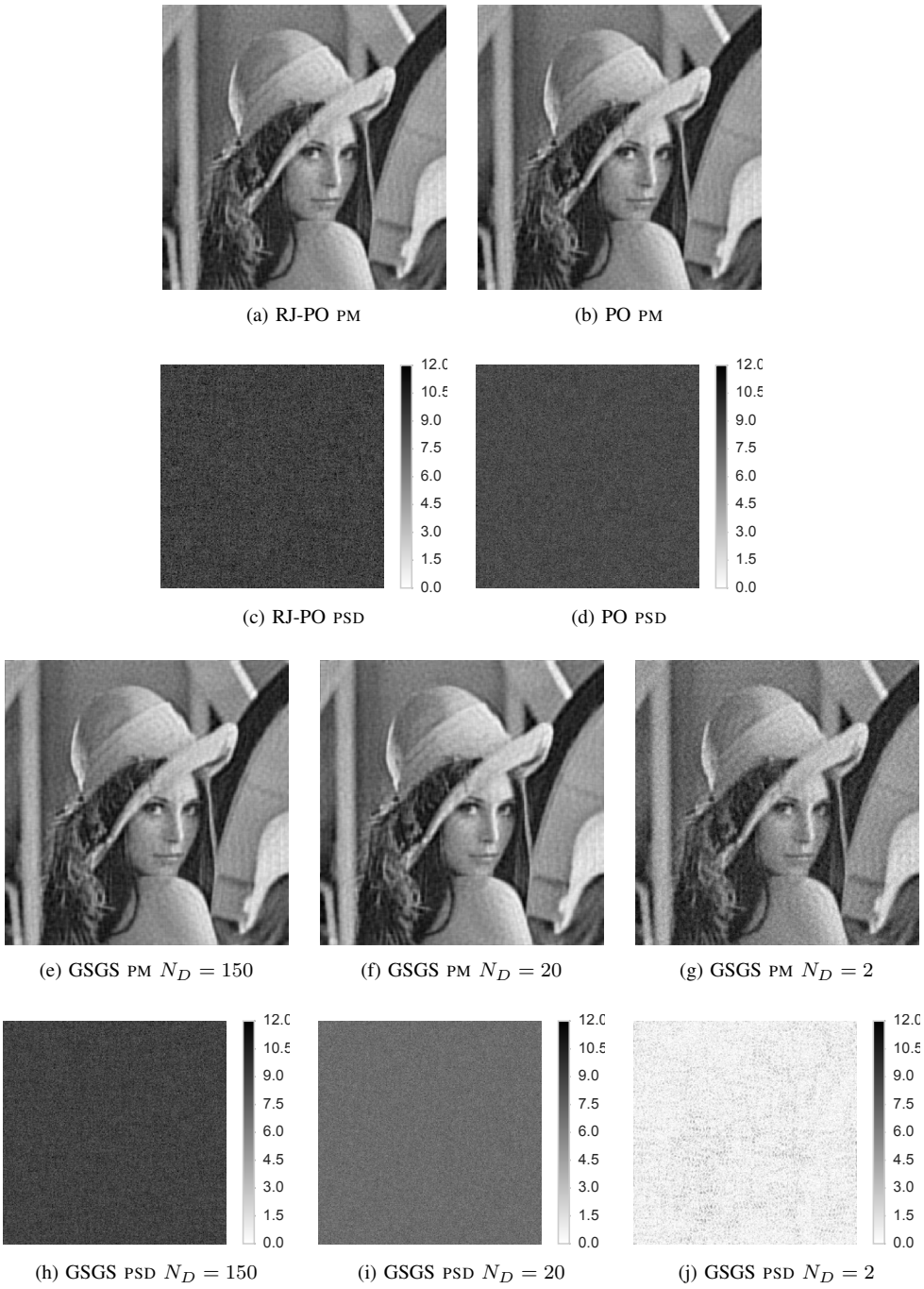
(c) $\gamma_{\mathbf{x}}$ as iteration (d) $\gamma_{\mathbf{x}}$ as time

Figure 2: Chains of hyper parameters $\gamma_{\mathbf{x}}$ and $\gamma_{\mathbf{n}}$.

To illustrate the effect of the perturbation for good space exploration, Fig. 3 shows the results when no perturbations $\boldsymbol{\varepsilon}^{(t)}$ are introduced and with $N_D = 10$. In this case, the hypotheses of Proposition 2 are no longer verified and those of Proposition 3 cannot be verified in practice. Moreover, the results show that both the covariance and the hyperparameters are wrongly estimated. This effect leads to an over-regularized image. A possible explanation is that the conjugate directions of the GSGS explore in a privileged way the directions of small variance (highest eigenvalues of $\mathbf{Q}$).

Regarding the computational cost, all the presented algorithms are dominated by the cost of the matrix-vector product $\mathbf{Qx}$. The cost thus depends on the specific problems and the structure of $\mathbf{Q}$ in the same way as for the conjugate gradient algorithm. For super-resolution problems, the cost of the matrix-vector product is almost equal to two discrete Fourier transforms of images. That said, the total number of matrix-vector products is related to N_D and the number of Gibbs iterations. Moreover, the computational cost is linear with respect to N_D.

The main concluding comment is that the proposed algorithm allows a great improvement in the convergence time of the GS. However the speed improvement can come with a bad covariance estimation if the number N_D of directions for the image $\mathbf{x}$ is not sufficient.



(a) RJ-PO PM (b) PO PM

(c) RJ-PO PSD (d) PO PSD

(e) GSGS PM $N_D = 150$ (f) GSGS PM $N_D = 20$ (g) GSGS PM $N_D = 2$

(h) GSGS PSD $N_D = 150$ (i) GSGS PSD $N_D = 20$ (j) GSGS PSD $N_D = 2$

Figure 1: Image results.

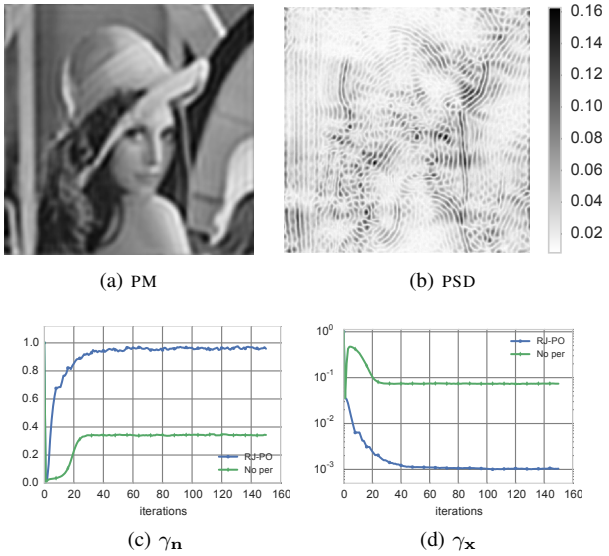
IV. Conclusion

The handling of high-dimensional distribution, especially Gaussian, appears in many linear inverse and estimation problems. With growing interest in "Big Data" and non stationary problems this task becomes critical. Moreover, the uncertainty around the estimated values, or the confidence interval, remains one of the difficult points combined with the hyperparameter estimation for automatic method designs.

The main contributions of this paper are (i) the proposition of a new algorithm in the class of the Gibbs samplers, able to address the case of high-dimensional Gaussian conditional distributions, and (ii) the convergence proof of the algorithm. It relies on a random excursion along a small set of directions instead of working with high dimensional distributions. The directions are appropriately chosen according to the gradient of the potential of the distribution.

This new algorithm is shown to be an efficient alternative to existing work like the PO-type algorithms: we ensure the theoretical convergence of the algorithm and, in some cases, we can show a drastic computing-time improvement.

The convergence of the algorithm is proved, provided



(a) PM (b) PSD

(c) $\gamma_{\mathbf{n}}$ (d) $\gamma_{\mathbf{x}}$

Figure 3: Results without perturbation and $N_D = 20$.

that a random perturbation around the gradient direction is introduced. Even if in theory the only condition to ensure convergence is to choose a perturbation distribution supported in the whole space, it appears in practice that the results are sensitive to the choice of the distribution. Moreover, the choice of the Gaussian distribution $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$ is the only case where the algorithm is more efficient than the PO and RJ-PO algorithms. The objective of further work will be to better understand this sensitivity and the open problem of the choice of the perturbation's distribution.

In further work the objective will be to study the convergence rate of the GSGS. In particular, the geometric ergodicity is an important property that ensures a fast convergence and allows us to give estimations of standard errors. The geometric ergodicity of Gibbs samplers has long been studied [43] and a lot of results are shown in the Gaussian case [44], as well as for applications in Bayesian hierarchical models [45], also in the case of joint Gaussian and Gamma distributions [46], [47], the latter being close to our illustration example.

Also, one has to choose the number N_D of mutually conjugate directions to sample at each iteration of the algorithm. In theory, this does not affect the convergence properties of the algorithm. As a perspective, one can propose an automatic choice of N_D, following the work in [40] for the RJ-PO. A research field could be the study of the algorithm's efficiency with respect to the eigenvalues of Q in the high dimensional case.

The proposed algorithm is somewhat independent of the chosen direction. The use of a preconditioner to compute direction, as in preconditioned conjugate gradient, should improve the computational cost by an N_D parameter smaller than at the present time. It depends, however, on each problem addressed.

From an experimental standpoint an additional assessment of the proposed method could rely on a numerical comparison with other existing approaches, for instance Hamltonian or Langevin algorithm [31]–[34].

This paper is focused on linear conditionally Gaussian models. By use of hidden variables, the algorithm should also be able to work with non Gaussian models that are still conditionally Gaussian.

APPENDIX

A. Proof of Proposition 2

This appendix is devoted to prove Proposition 2. It is mainly inspired by the proofs presented in [28] (see also [27], [29]) for different random scan strategies in order to sample $p(\mathbf{x}|\boldsymbol{\theta})$. The only difference is that the random choice is not according to a set of coordinates of $\mathbf{x}$ in the canonical basis, but according to a mutually conjugate set with respect to a current matrix $\mathbf{Q}_{\boldsymbol{\theta}}$. Therefore the same arguments as detailed in [28] can be used to prove the irreducibility: if the support of the density $p(\boldsymbol{\varepsilon})$ is $\mathbb{R}^N$, all the directions can be explored in one step of the algorithm. Therefore any $\mathbf{y} \in \mathbb{R}^N$ can be reached in one step by taking, for example, $\mathbf{d}_1 = \mathbf{x}^{(t-1)} - \mathbf{y}$, $\widetilde{\alpha}_1 = 1$, $\widetilde{\alpha}_n = 0$, $n = 2, \ldots, N_D$. Using classical continuity arguments, we can deduce that the probability of reaching any open ball $\mathcal{B}(\mathbf{y}, r)$, centered in $\mathbf{y}$ of radius r, conditional to any current point $\mathbf{x}^{(t)}$, is strictly positive, which ensures the chain to be irreducible.

The rest of the proof focuses on the fact that $p(\mathbf{x}, \boldsymbol{\theta})$ is an invariant probability of the chain. We use the same arguments and notations of [28]. Let $\mathbf{x} \in \mathbb{R}^N$ and a set D of mutually conjugate directions with respect to a definite positive matrix $\mathbf{Q}$. We decompose $\mathbf{x} = (\mathbf{x}_D, \mathbf{x}_{\setminus D})$ which is always possible as explained in section II-A.

Define $(\mathbf{x}', \boldsymbol{\theta}') \in \mathbb{R}^N \times \Theta$ a current point and $(\mathbf{x}', \boldsymbol{\theta}') \in \mathbb{R}^N \times \Theta$ the point obtained by Algorithm 1 with the transition Kernel:

$$P(\mathbf{x}, \boldsymbol{\theta}|\mathbf{x}', \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}|\mathbf{x}', \boldsymbol{\theta}')\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D}, \mathbf{x}', \boldsymbol{\theta})\delta(\mathbf{x}_{\setminus D} - \mathbf{x}'_{\setminus D})$$

with π denoting any conditional probability and δ is the Dirac function. The objective is to show that if $(\mathbf{x}', \boldsymbol{\theta}')$ is distributed according to the joint distribution p, then $(\mathbf{x}, \boldsymbol{\theta})$ is also distributed according to p.

Let $A \subset \mathbb{R}^N$ be a measurable set. The following lines are the result of the definition of the transition Kernel, the use of the general product rule, and of sequential integration with

respect to $\boldsymbol{\theta}'$, $\mathbf{x}'_D$ and $\mathbf{x}'_{\setminus D}$:

$$\begin{aligned}
&\mathbb{P}((\mathbf{x},\boldsymbol{\theta})\in A)\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta})P(\mathbf{x},\boldsymbol{\theta}|\mathbf{x}',\boldsymbol{\theta}')p(\mathbf{x}',\boldsymbol{\theta}')\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'\mathrm{d}\boldsymbol{\theta}'\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}',\boldsymbol{\theta}')\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\mathbf{x}',\boldsymbol{\theta})\ldots\\
&\qquad \ldots\delta(\mathbf{x}_{\setminus D}-\mathbf{x}'_{\setminus D})p(\mathbf{x}',\boldsymbol{\theta}')\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'\mathrm{d}\boldsymbol{\theta}'\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta})p(\mathbf{x}',\boldsymbol{\theta})\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\mathbf{x}',\boldsymbol{\theta})\ldots\\
&\qquad \ldots\delta(\mathbf{x}_{\setminus D}-\mathbf{x}'_{\setminus D})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta})p(\mathbf{x}'_{\setminus D},\boldsymbol{\theta})\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\mathbf{x}'_{\setminus D},\boldsymbol{\theta})\ldots\\
&\qquad \ldots\delta(\mathbf{x}_{\setminus D}-\mathbf{x}'_{\setminus D})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'_{\setminus D}\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta})p(\mathbf{x}_{\setminus D},\boldsymbol{\theta})\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\boldsymbol{\theta})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta})p(\mathbf{x},\boldsymbol{\theta})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

Hence the joint probability $p(\mathbf{x},\boldsymbol{\theta})$ is an invariant probability of the Markov chain produced by Algorithm 1.

B. Proof of Proposition 3

This appendix is dedicated to prove Proposition 3. Let $(\mathbf{x}^{(0)},\boldsymbol{\theta}^{(0)})\in\mathbb{R}^N\times\Theta$ be a current point and $(\mathbf{x}^{(t)},\boldsymbol{\theta}^{(t)})$ the point produced by the chain of Algorithm 1 at iteration t. The objective is to prove that for any non-negligible subset $A\subset\mathbb{R}^N\times\Theta$, there is $T\geq 0$ such as $\mathbb{P}((\mathbf{x}^{(T)},\boldsymbol{\theta}^{(T)})\in A|\mathbf{x}^{(0)},\boldsymbol{\theta}^{(0)})>0$. Using the hypothesis H-2, it is sufficient to prove that for any non-negligible subset $A_x\in\mathbb{R}^N$, there is $T\geq 0$ such as:

$$\mathbb{P}(\mathbf{x}^{(T)}\in A_x|\mathbf{x}^{(0)},\boldsymbol{\theta}^{(0)})>0 \tag{5}$$

Given $\mathbf{x}^{(0)}$, we denote by $\boldsymbol{\theta}$ the corresponding element that respects conditions H-3. It is sufficient to prove the Proposition in the following framework:
F-1 $\boldsymbol{\theta}^{(N+1)}=\boldsymbol{\theta}^{(N)}=\ldots=\boldsymbol{\theta}^{(0)}=\boldsymbol{\theta}$,
F-2 $\mathbf{m}_{\boldsymbol{\theta}}=0$,
F-3 $\mathbf{Q}_{\boldsymbol{\theta}}=\mathrm{diag}(q_1,\ldots,q_N)$ is diagonal.
Indeed, if we prove the inequality (5) with fixed $\boldsymbol{\theta}$ for $N+1$ iterations, continuity arguments using conditions H-1 and H-2 will end the proof of the Proposition. The simplifications F-2 and F-3 can be assumed by a change of variable $\mathbf{y}^{(t)}=\mathbf{x}^{(t)}-\mathbf{m}_{\boldsymbol{\theta}}$ and by considering the basis of $\mathbb{R}^N$ formed by the eigenvectors of $\mathbf{Q}_{\boldsymbol{\theta}}$.

In this simplified framework, the chain of Algorithm 1 produces $\mathbf{x}^{(t)},\ t=1,\ldots,N+1$, such as:

$$\mathbf{x}^{(t)}=(\mathbf{I}-\alpha^{(t)}\mathbf{Q}_{\boldsymbol{\theta}})(\mathbf{I}-\alpha^{(t-1)}\mathbf{Q}_{\boldsymbol{\theta}})\ldots(\mathbf{I}-\alpha^{(1)}\mathbf{Q}_{\boldsymbol{\theta}})\mathbf{x}^{(0)},$$

with $\mathbf{I}$ the identity matrix in $\mathbb{R}^N$ and, noting $\mathbf{x}=(x_1,\ldots,x_N)^{\mathrm{t}}$, we have, for $n=1,\ldots,N$:

$$\mathbf{x}_n^{(t)}=(1-\alpha^{(t)}q_n)(1-\alpha^{(t-1)}q_n)\ldots(1-\alpha^{(1)}q_n)\mathbf{x}_n^{(0)}. \tag{6}$$

The hypothesis H-3.2 ensures that $\mathbf{x}_n^{(0)}\neq 0,\ n=1,\ldots,N$, therefore we can assume without loss of generality that $\mathbf{x}_n^{(0)}=1,\ n=1,\ldots,N$, and equation (6) is, in this case:

$$\mathbf{x}_n^{(t)}=(1-\alpha^{(t)}q_n)(1-\alpha^{(t-1)}q_n)\ldots(1-\alpha^{(1)}q_n). \tag{7}$$

The following Lemma proves that any point in $\mathbb{R}^N$ can be reached by the chain in $N+1$ iterations.

Lemma 1. *For any $\mathbf{y}\in\mathbb{R}^N$, there is $\alpha=(\alpha^{(1)},\ldots,\alpha^{(N+1)})$ such as $\mathbf{x}^{(N+1)}=\mathbf{y}$, where $\mathbf{x}^{(N+1)}$ is defined by* (7) *with* $t=N+1$.

Proof. This can be done by interpreting it as an interpolation problem: given $\mathbf{y}\in\mathbb{R}^N$, the objective is to show that there is a polynomial P_α^{N+1} such as:

$$\begin{aligned}
P_\alpha^{N+1}(q_n) &= y_n,\quad n=1,\ldots,N &(8)\\
P_\alpha^{N+1}(0) &= 1 &(9)
\end{aligned}$$

with P_α^{N+1} defined by the right hand side of (7) with $t=N+1$. The constraint (9) is due to the specific form of P_α^{N+1}. Also the fact that the parameters $\alpha^{(n)}$ must be real, implies that the polynomial P_α^{N+1} must have only real roots. It is well known that there is a polynomial of degree N that respects (8) and (9). Let us denote by Q such a polynomial. But the roots of Q may be complex. However we can show that there is a polynomial of degree $N+1$ with real roots that respects the conditions (8) and (9). Indeed, let us consider the polynomial Q and a polynomial R of degree $N+1$ such as $R(q_1)=R(q_2)=\ldots=R(q_N)=R(0)=0$. Therefore any polynomial $P_\tau=Q+\tau R$, $\tau\in\mathbb{R}$, respects conditions (8) and (9), and it is trivial to show that for τ^* sufficiently large, the polynomial P_{τ^*} has all its roots $r_n^*\in\mathbb{R},\ n=1,\ldots,N$. Therefore, taking $P_\alpha^{N+1}=P_{\tau^*}$, *i.e.* $\alpha^{(n)}=1/r_n^*$ ends the proof of the lemma. □

Using this lemma and the continuity of P_α^{N+1} with respect to α, it is trivial to prove (5) and then the Proposition.

REFERENCES

[1] J. Idier, Ed., *Bayesian Approach to Inverse Problems*. London: ISTE Ltd and John Wiley & Sons Inc., 2008.

[2] J.-F. Giovannelli and J. Idier, Eds., *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*. London: ISTE Ltd and John Wiley & Sons Inc., 2015.

[3] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[4] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 932–946, July 1995.

[5] J.-F. Giovannelli, "Unsupervised Bayesian convex deconvolution based on a field with an explicit partition function," *IEEE Trans. Image Processing*, vol. 17, no. 1, pp. 16–26, Jan. 2008.

[6] X. Tan, J. Li, and P. Stoica, "Efficient sparse Bayesian learning via Gibbs sampling," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 3634 –3637.

[7] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "Blind deconvolution of sparse pulse sequences under a minimum distance constraint: a partially collapsed Gibbs sampler method," *IEEE Trans. Signal Processing*, vol. 60, no. 6, pp. 2727–2743, June 2012.

[8] O. Féron, B. Duchêne, and A. Mohammad-Djafari, "Microwave imaging of piecewise constant objects in a 2D-TE configuration," *International Journal of Applied Electromagnetics and Mechanics*, vol. 26, no. 6, pp. 167–174, IOS Press 2007.

[9] H. Ayasso and A. Mohammad-Djafari, "Joint NDT image restoration and segmentation using Gauss-Markov-Potts prior models and variational Bayesian computation," *IEEE Trans. Image Processing*, vol. 19, no. 9, pp. 2265–2277, 2010.

[10] F. Orieux, J.-F. Giovannelli, and T. Rodet, "Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution," *J. Opt. Soc. Amer.*, vol. 27, no. 7, pp. 1593–1607, July 2010.

[11] Q. Liu, E. Chen, B. Xiang, C. H. Q. Ding, and L. He, "Gaussian process for recommender systems," in *Lecture Notes in Computer Science, Knowledge Science, Engineering and Management*, vol. 7091, 2011, pp. 56–67.

[12] J. E. Besag, "On the correlation structure of some two-dimentional stationary processes," *Biometrika*, vol. 59, no. 1, pp. 43–48, 1972.

[13] H. Rue, "Fast sampling of Gaussian Markov random fields," *J. R. Statist. Soc. B*, vol. 63, no. 2, pp. 325–338, 2001.

[14] A. E. Gelfand, H.-J. Kim, C. F. Sirmans, and S. Banerjee, "Spatial modeling with spatially varying coefficient processes," *J. Amer. Statist. Assoc.*, vol. 98, no. 462, pp. 387–396, 2003.

[15] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 33, no. 4, pp. 959–963, Aug. 1985.

[16] R. Chellappa and A. Jain, *Markov Random Fields: Theory and Application*. Academic Press Inc, 1992.

[17] J. M. Bardsley, "MCMC-based image reconstruction with uncertainty quantification," *SIAM Journal of Scientific Computation*, vol. 34, no. 3, pp. A1316–A1332, 2012.

[18] J. M. Bardsley, M. Howard, and J. G. Nagy, "Efficient MCMC-based image deblurring with Neumann boundary conditions," *Electronic Transactions on Numerical Analysis*, vol. 40, pp. 476–488, 2013.

[19] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, ser. Monographs on Statistics and Applied Probability. Chapman & Hall, 2005, vol. 104.

[20] P. Lalanne, D. Prévost, and P. Chavel, "Stochastic artificial retinas: algorithm, optoelectronic circuits, and implementation," *Applied Optics*, vol. 40, no. 23, pp. 3861–3876, 2001.

[21] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer Verlag, Berlin Germany, 2003.

[22] G. Papandreou and A. Yuille, "Gaussian sampling by local perturbations," in *Proc. Int. Conf. on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec. 2010, pp. 1858–1866.

[23] F. Orieux, J.-F. Giovannelli, T. Rodet, H. Ayasso, and A. Abergel, "Super-resolution in map-making based on a physical instrument model and regularized inversion. Application to SPIRE/Herschel." *Astron. Astrophys.*, vol. 539, Mar. 2012.

[24] C. Gilavert, S. Moussaoui, and J. Idier, "Rééchantillonnage gaussien en grande dimension pour les problèmes inverses," in *Actes 24^{e} coll. GRETSI*, Brest, France, Sep. 2013.

[25] C. Fox, "A conjugate direction sampler for normal distributions with a few computed examples," Electronics Technical Report No. 2008-1, University of Otago, Dunedin, New Zealand, Tech. Rep., 2008.

[26] A. Parker and C. Fox, "Sampling Gaussian distributions in Krylov spaces with conjugate gradients," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. B312–B334, 2012.

[27] R. A. Levine, Z. Yu, W. G. Hanley, and J. J. Nitao, "Implementing random scan Gibbs samplers," *Computational Statistics*, vol. 20, no. 1, pp. 177–196, 2005.

[28] R. A. Levine and G. Casella, "Optimizing random scan Gibbs samplers," *Journal of Multivariate Analysis*, vol. 97, no. 10, pp. 2071–2100, 2006.

[29] K. Latuszynski, G. O. Roberts, and J. Rosenthal, "Adaptive Gibbs samplers and related MCMC methods," *Annals of Applied Probability*, vol. 23, no. 1, pp. 66–98, 2013.

[30] F. Orieux, O. Féron, and J.-F. Giovannelli, "Sampling high-dimensional Gaussian fields for general linear inverse problem," *IEEE Signal Proc. Let.*, vol. 19, no. 5, pp. 251–254, May 2012.

[31] O. Stramer and R. L. Tweedie, "Langevin-type models i: Diffusions with given stationary distributions, and their discretizations," *Methodology and Computing in Applied Probability*, vol. 1, no. 3, pp. 283–306, 1999.

[32] ——, "Langevin-type models ii: Self-targeting candidates for MCMC algorithms," *Methodology and Computing in Applied Probability*, vol. 1, no. 3, pp. 307–328, 1999.

[33] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216–222, 1987.

[34] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, G. J. S. Brooks, A. Gelman and X.-L. Meng, Eds. Chapman & Hall, 2010, ch. 5, pp. 113–162.

[35] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu, "Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 3964–3967.

[36] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in practice*. Boca Raton, USA: Chapman & Hall/CRC, 1996.

[37] G. O. Roberts and S. Rosenthal, "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains," *The Annals of Applied Probability*, vol. 16, no. 4, pp. 2123–2139, 2006.

[38] S. Meyn and R. Tweedie, *Markov chains and stochastic stability*. Springer-Verlag, London, 1993.

[39] F. Orieux, O. Féron, and J.-F. Giovannelli, "Gradient scan Gibbs sampler: An efficient high-dimensional sampler. Application in inverse problems," in *ICASSP*, Apr. 2015.

[40] C. Gilavert, S. Moussaoui, and J. Idier, "Efficient Gaussian sampling for solving large-scale inverse problems using MCMC," *IEEE Trans. Image Processing*, vol. 63, no. 1, pp. 70–80, Jan. 2015.

[41] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Proc. Mag.*, pp. 21–36, May 2003.

[42] G. Rochefort, F. Champagnat, G. Le Besnerais, and J.-F. Giovannelli, "An improved observation model for super-resolution under affine motion," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3325–3337, Nov. 2006.

[43] G. O. Roberts and N. G. Polson, "On the geometric convergence of the Gibbs sampler," *J. R. Statist. Soc. B*, vol. 56, no. 2, pp. 377–384, 1994.

[44] G. O. Roberts and S. K. Sahu, "updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler," *J. R. Statist. Soc. B*, vol. 59, no. 2, pp. 291–317, 1997.

[45] P. O. and G. O. Roberts, "Stability of the Gibbs sampler for Bayesian hierarchical models," *Ann. Statist.*, vol. 36, no. 1, pp. 95–117, 2008.

[46] J. P. Hobert and C. J. Geyer, "Geometric ergodicity of Gibbs and block samplers for a hierarchical random effects model," *Journal of Multivariate Analysis*, vol. 67, no. 2, pp. 414–430, 1998.

[47] A. Johnson and O. Burbank, "Geometric ergodicity and scanning strategies for two-component Gibbs samplers," *Communications in Statistics - Theory and Methods*, vol. 44, no. 15, pp. 3125–3145, 2015.