



HAL
open science

Regularization by denoising: Bayesian model and Langevin-within-split Gibbs sampling

Elhadji C. Faye, Mame Diarra Fall, Nicolas Dobigeon

► **To cite this version:**

Elhadji C. Faye, Mame Diarra Fall, Nicolas Dobigeon. Regularization by denoising: Bayesian model and Langevin-within-split Gibbs sampling. 2024. hal-04480090

HAL Id: hal-04480090

<https://hal.science/hal-04480090v1>

Preprint submitted on 27 Feb 2024

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

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

Regularization by denoising: Bayesian model and Langevin-within-split Gibbs sampling

Elhadji C. Faye, Mame Diarra Fall and Nicolas Dobigeon, *Senior Member, IEEE*

Abstract—This paper introduces a Bayesian framework for image inversion by deriving a probabilistic counterpart to the regularization-by-denoising (RED) paradigm. It additionally implements a Monte Carlo algorithm specifically tailored for sampling from the resulting posterior distribution, based on an asymptotically exact data augmentation (AXDA). The proposed algorithm is an approximate instance of split Gibbs sampling (SGS) which embeds one Langevin Monte Carlo step. The proposed method is applied to common imaging tasks such as deblurring, inpainting and super-resolution, demonstrating its efficacy through extensive numerical experiments. These contributions advance Bayesian inference in imaging by leveraging data-driven regularization strategies within a probabilistic framework.

I. INTRODUCTION

This paper is interested in conducting Bayesian inference about an image $\mathbf{x} \in \mathbb{R}^n$ given the measurements $\mathbf{y} \in \mathbb{R}^m$ related to \mathbf{x} through a statistical model specified by the likelihood function

$$p(\mathbf{y}|\mathbf{x}) \propto \exp[-f(\mathbf{x}, \mathbf{y})]. \quad (1)$$

In (1), the potential function $f(\mathbf{x}, \mathbf{y})$ is a fidelity term, i.e., accounting for the consistency of \mathbf{x} with respect to (w.r.t.) the measured data \mathbf{y} . In what follows, this potential function will be assumed to be convex and L_f -smooth, i.e., continuously differentiable and its gradient is Lipschitz continuous with Lipschitz constant L_f . This problem is in line with the most frequently encountered imaging inverse problems such as denoising, deblurring, and inpainting relying on a linear forward model and a Gaussian perturbation. For such tasks, the potential function writes $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2\sigma^2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the degradation matrix. Estimating \mathbf{x} from \mathbf{y} is generally an ill-posed or, at least, ill-conditioned problem. The Bayesian paradigm consists in assigning a prior distribution to \mathbf{x} , which summarizes the prior knowledge about \mathbf{x} and acts as a regularization. This prior distribution writes

$$p(\mathbf{x}) \propto \exp[-\beta g(\mathbf{x})] \quad (2)$$

where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ stands for the regularization term and the parameter $\beta > 0$ controls the amount of regularization enforced

by the prior distribution. The posterior distribution $p(\mathbf{x}|\mathbf{y})$ is derived from the likelihood $p(\mathbf{y}|\mathbf{x})$ and the prior distribution $p(\mathbf{x})$ using the Bayes' rule

$$p(\mathbf{x}|\mathbf{y}) \propto \exp[-f(\mathbf{x}, \mathbf{y}) - \beta g(\mathbf{x})]. \quad (3)$$

This posterior distribution provides a comprehensive description of the solutions and allows various Bayesian estimators and uncertainty measures to be derived. In particular, computing the maximum a posteriori (MAP) estimator boils down to solving the minimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) + \beta g(\mathbf{x}). \quad (4)$$

Numerous works from the literature have focused on the difficult task of designing a relevant prior distribution $p(\mathbf{x})$ or, equivalently, a relevant potential function $g(\mathbf{x})$. These regularizations usually promote specific expected or desired properties about \mathbf{x} . More specifically, conventional optimization methods solving (4) are generally based on explicit model-based regularizations, such as total variation (TV) promoting piecewise constant behavior [1], Sobolev promoting smooth content [2] or sparsity-promoting regularizations based on the use of ℓ_p -norm with $p \leq 1$ [3], [4]. However, designing an appropriate model-based regularization remains an empirical and subjective choice. Moreover, their ability of characterizing complex image structures is generally limited or comes at the price of a significant increase of the resulting algorithmic burden. More recently, a different route has been taken by devising smart strategies avoiding the ad hoc design of explicit model-based regularizations. The seminal work by Venkatakrisnan *et al.* has introduced the concept of *plug-and-play* (PnP) as an implicit prior [5]. This framework naturally emerges whenever the algorithmic scheme designed to solve (4) embeds the proximal operator associated with $g(\cdot)$. Possible schemes include the alternating direction method of multipliers (ADMM) [4], half quadratic splitting (HQS) [6] or Douglas-Rachford splitting [7]. Interestingly, this proximal mapping can be interpreted as a denoising task under the assumption of an additive white Gaussian noise. PnP approaches replace this proximal step by a more general denoiser $D : \mathbb{R}^n \rightarrow \mathbb{R}^n$, including non-local means (NLM) [8], block-matching and 3D filtering (BM3D) [9] or any more recently proposed learning-based denoisers such as DnCNN [10] or DRUNet [11]. Thanks to its effectiveness and its simplicity, this framework has gained in popularity for a wide range of applications in the context of imaging problems [11]–[14]. In the same vein as PnP, the regularization-by-denoising (RED) framework defines an explicit image-adaptive Laplacian-based prior which only relies

E. C. Faye and M. D. Fall are with Institut Denis Poisson, University of Orléans, Orléans, France (e-mail: {elhadji-cisse.faye, diarra.fall}@univ-orleans.fr).

N. Dobigeon is with University of Toulouse, IRIT/INP-ENSEEIH, CNRS, 2 rue Charles Camichel, BP 7122, 31071 Toulouse Cedex 7, France (e-mail: Nicolas.Dobigeon@enseeiht.fr).

Part of this work was supported by the Artificial Natural Intelligence Toulouse Institute (ANITI, ANR-19-PI3A-0004), the ALiO Project (ANR-20-THIA-0017) and the BACKUP project (ANR-23-CE40-0018-01).

on the ability of performing a denoising task [15]. Empirically, RED has shown to outperform PnP-based approaches and has motivated several subsequent research works [14], [16], [17].

All the variational approaches discussed above treat \mathbf{x} in a deterministic way and generally produce only point estimates approximating the solution of the minimization problem (4). As an alternative, the Bayesian framework models the image \mathbf{x} as a random variable and generally seeks a comprehensive description of the posterior distribution $p(\mathbf{x}|\mathbf{y})$. As such, Bayesian methods are able to go beyond a sole point estimation, by enhancing it with a quantification of uncertainty in a probabilistic manner in terms of variance and credibility intervals. This ability to quantify uncertainty is particularly useful for decision-making and reliability assessment [18], [19]. Exploring the posterior distribution is generally carried out by generating samples asymptotically distributed according to this target distribution using Markov chain Monte Carlo (MCMC) methods. Most of the works dedicated to the development of MCMC algorithms for inverse problems in imaging relies on conventional model-based prior distributions. As their deterministic counterparts, they encode expected characteristics of the image prescribed beforehand and chosen based on quite empirical arguments. Very few recent works have attempted to depart from this paradigm by incorporating data- or task-driven regularizations as prior distributions. For instance, available training samples can be used to learn a mapping from an instrumental latent distribution towards the image prior. Benefiting from advances in the machine learning literature, this mapping can be chosen as a deep generative model, such as a variational autoencoder [20] or a normalizing flow [21]. Devising a PnP prior in the context of Monte Carlo sampling has been investigated in [22], resulting in the so-called PnP unadjusted Langevin algorithm (PnP-ULA). Its rationale follows the same motivation as its deterministic counterpart, namely avoiding the explicit definition of the prior distribution by the ability of performing a denoising task. Surprisingly, although that the RED approach has shown to outperform PnP when embedded into a variational framework, up to authors' knowledge, there is no equivalent for the RED paradigm, i.e., RED has never been formulated into a Bayesian framework and embedded into a Monte Carlo algorithm.

The objective of this paper is to fill this gap. More precisely, the main contributions reported hereafter can be summarized as follows. First, Section II introduces a probabilistic counterpart of RED by defining a new distribution that can be subsequently chosen as a prior distribution in a Bayesian inversion task. Then, Section III introduces a new Monte Carlo algorithm that is shown to be particularly well suited to sample from the resulting posterior distribution. It follows an asymptotically exact data augmentation (AXDA) scheme [23], resulting in a nonstandard instance of the split Gibbs sampler (SGS) [24]. This sampling scheme is thus accompanied by a thorough theoretical analysis to ensure and quantify its convergence. The rationale of the proposed approach is also put into perspective w.r.t. recently proposed Monte Carlo algorithms, in particular PnP-ULA, drawing some connections between AXDA and RED leveraging the Tweedie's formula [25]. Then, the proposed algorithm is instantiated to solve three

ubiquitous inversion tasks, namely deblurring, inpainting and super-resolution. Extensive numerical experiments are conducted in Section IV to compare the performance of the proposed algorithm to state-of-the-art variational and Monte Carlo methods.

Notations and conventions. The Euclidean norm on \mathbb{R}^n is denoted by $\|\cdot\|$. We denote by $\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and precision matrix \mathbf{Q} . The $(n \times n)$ -identity matrix is denoted \mathbf{I}_n . For any matrix $\mathbf{S} \in \mathcal{M}_n(\mathbb{R})$, if we denote $\mathbf{0}$ the zero matrix, the notation $\mathbf{0} \preceq \mathbf{S}$ means that \mathbf{S} is semi-definitive positive. The Wasserstein distance of order 2 between two probability measures τ and τ' on \mathbb{R}^n with finite 2-moments is defined by $W_2(\tau, \tau') = (\inf_{\zeta \in \mathcal{T}(\tau, \tau')} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|V - V'\|^2 d\zeta(V, V'))^{1/2}$, where $\mathcal{T}(\tau, \tau')$ is the set of transport plans of τ and τ' .

II. BAYESIAN FORMULATION OF RED INVERSION

This section starts by recalling some background about RED. Then it proposes a probabilistic counterpart of the regularization, that can be subsequently used as a prior within a Bayesian framework.

A. Regularization by denoising (RED)

The RED engine defines $g(\cdot)$ as the explicit image-adaptive Laplacian-based potential [15]

$$g_{\text{red}}(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top (\mathbf{x} - \mathbf{D}_\nu(\mathbf{x})) \quad (5)$$

where $\mathbf{D}_\nu : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a denoiser with ν controlling the denoising strength, designed for the removal of additive white Gaussian noise. Although it offers a significant flexibility in the choice of the denoisers that can be used, RED requires $\mathbf{D}_\nu(\cdot)$ to obey the following assumptions, referred to as the RED conditions.

(C1) *Local homogeneity*: $\forall \mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{D}_\nu((1 + \epsilon)\mathbf{x}) = (1 + \epsilon)\mathbf{D}_\nu(\mathbf{x}) \quad (6)$$

for any sufficiently small $\epsilon > 0$.

(C2) *Differentiability*: the denoiser $\mathbf{D}_\nu(\cdot)$ is differentiable with Jacobian denoted $\nabla \mathbf{D}_\nu(\cdot)$.

(C3) *Jacobian symmetry* [16]: $\forall \mathbf{x} \in \mathbb{R}^n$, $\nabla \mathbf{D}_\nu(\mathbf{x})^\top = \nabla \mathbf{D}_\nu(\mathbf{x})$.

(C4) *Strong passivity*: the Jacobian spectral radius satisfies $\eta(\nabla \mathbf{D}_\nu(\mathbf{x})) \leq 1$.

The major implication of local homogeneity (C1) is that the directional derivative of $\mathbf{D}_\nu(\cdot)$ along \mathbf{x} can be computed by applying the denoiser itself, i.e.,

$$\nabla \mathbf{D}_\nu(\mathbf{x})\mathbf{x} = \mathbf{D}_\nu(\mathbf{x}). \quad (7)$$

The Jacobian symmetry (C3) and the strong passivity (C4) ensure that applying the denoiser does not increase the norm of the input:

$$\|\mathbf{D}_\nu(\mathbf{x})\| = \|\nabla \mathbf{D}_\nu(\mathbf{x})\mathbf{x}\| \leq \eta(\nabla \mathbf{D}_\nu(\mathbf{x})) \cdot \|\mathbf{x}\| \leq \|\mathbf{x}\|. \quad (8)$$

Interestingly, two additional keys and highly beneficial properties follow: *i)* the RED potential $g_{\text{red}}(\cdot)$ is a convex functional

and *ii*) the gradient of $g_{\text{red}}(\cdot)$ is expressed as the denoising residual

$$\nabla g_{\text{red}}(\mathbf{x}) = R(\mathbf{x}) = \mathbf{x} - D_\nu(\mathbf{x}) \quad (9)$$

which avoids differentiating the denoising operation itself. Thus, one of the most appealing opportunity offered by RED is its ability to embed powerful denoisers, such as those based on deep neural networks, without requiring to differentiate them. It is worth noting that if the denoising function $D_\nu(\cdot)$ does not meet the condition (C3), then there is no regularizer $g(\cdot)$ whose gradient can be written as the residual $R(\cdot)$ [16]. Unfortunately many popular denoisers, such as trainable nonlinear reaction-diffusion (TNRD), NLM, BM3D, and DnCNN, are characterized by non-symmetric Jacobian. Yet, RED-based restoration algorithms are shown to empirically converge and to reach excellent performance when solving various inverse problems even when those conditions are partially satisfied [15], [17].

B. Probabilistic counterpart of RED

To formulate the RED-based inversion within a statistical framework, one requirement consists in introducing a prior distribution defined from the RED potential $g_{\text{red}}(\cdot)$ given by (5). More precisely, one defines

$$p_{\text{red}}(\mathbf{x}) \propto \exp \left[-\frac{\beta}{2} \mathbf{x}^\top (\mathbf{x} - D_\nu(\mathbf{x})) \right]. \quad (10)$$

The functional $p_{\text{red}}(\cdot)$ does not necessarily define a probability density function (pdf). For $p_{\text{red}}(\cdot)$ to be a valid pdf, i.e. $\int_{\mathbb{R}^n} p_{\text{red}}(\mathbf{x}) d\mathbf{x} < \infty$, certain conditions must be satisfied.

Assumption 1. *The matrix $\Lambda(\mathbf{x}) = \mathbf{I}_n - \nabla D_\nu(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^n$, has at least one non-zero eigenvalue.*

This technical assumption is not restrictive and it is easy to show that it would be violated only in trivial cases. Indeed, let $\bar{\mathbf{x}} \in \mathbb{R}^n$ denote an image such that all the eigenvalues of $\Lambda(\bar{\mathbf{x}})$ are zero. Then the matrix $\nabla D_\nu(\bar{\mathbf{x}})$ is symmetric (see RED condition (C3)), with real coefficients and all eigenvalues equal to 1. The spectral theorem yields $\nabla D_\nu(\bar{\mathbf{x}}) = \mathbf{I}_n$. From the local homogeneity (C1) and its corollary (7), one has $D_\nu(\bar{\mathbf{x}}) = \bar{\mathbf{x}}$, i.e., $\bar{\mathbf{x}}$ is already a noise-free image which does not need to be further denoised.

The next result states that mild assumptions are sufficient to guarantee that the function (10) defines a proper distribution.

Proposition 1. *If Assumption 1 and Conditions (C3)–(C4) hold, then*

$$\int_{\mathbb{R}^n} p_{\text{red}}(\mathbf{x}) d\mathbf{x} < +\infty \quad (11)$$

and $p_{\text{red}}(\cdot)$ in (10) defines a proper pdf.

Proof. See Appendix A. \square

C. RED posterior distribution

Combining the RED prior $p_{\text{red}}(\mathbf{x})$ defined by (10) and the likelihood function $p(\mathbf{y}|\mathbf{x})$ defined by (1), the RED posterior distribution π of interest writes

$$\begin{aligned} \pi(\mathbf{x}) &\triangleq p(\mathbf{x}|\mathbf{y}) \\ &\propto \exp \left[-f(\mathbf{x}, \mathbf{y}) - \frac{\beta}{2} \mathbf{x}^\top (\mathbf{x} - D_\nu(\mathbf{x})) \right]. \end{aligned} \quad (12)$$

As stated earlier, deriving the MAP estimator associated with the RED posterior (12) consists in solving the optimization problem (4). In the seminal paper [15], this problem is tackled thanks to first-order optimization methods such as steepest decent (SD), fixed-point (FP) iteration and ADMM. More recently, it has been reformulated as a convex optimization problem using a projection onto the fixed point set of demi-contractive denoisers [17]. Instead, the work conducted in this manuscript proposes to follow a different route by proposing to sample from this posterior distribution. While these samples offer a comprehensive description of the RED posterior, they can be subsequently used to derive Bayesian estimators or credibility intervals. Because of the non-standard form of the RED posterior, sampling according to (12) requires to develop a dedicated algorithm introduced in the following section.

III. PROPOSED ALGORITHM

A. Langevin-within-split Gibbs sampler

Generating samples efficiently from the posterior distribution with pdf $\pi(\mathbf{x})$ defined by (12) is not straightforward, in particular due to the use of the denoiser $D_\nu(\cdot)$. When $\pi(\cdot)$ is proper and smooth with $\mathbf{x} \mapsto \nabla \log \pi(\mathbf{x})$ Lipschitz continuous, one solution would consist in resorting to the ULA [26]. This strategy will be shown to be intimately related to PnP-ULA in Section III-D. However, it may suffer from several shortcomings, such as poor mixing properties and higher resulting computational times (see experimental results in Section IV-D). Conversely, the work in this manuscript derives a dedicated Monte Carlo algorithm to sample from a posterior distribution written as (3). This algorithm will be shown to be particularly well suited to sample from the RED posterior (12), i.e., when $g(\cdot) = g_{\text{red}}(\cdot)$.

The proposed sampling scheme first leverages an asymptotically exact data augmentation (AXDA) as introduced in [23]. Inspired by optimization-flavored counterparts, AXDA employs a variable splitting technique to simplify and speed up the sampling according to possibly complex distributions. More precisely, it introduces an auxiliary variable $\mathbf{z} \in \mathbb{R}^n$ and considers the augmented distribution

$$\begin{aligned} \pi_\rho(\mathbf{x}, \mathbf{z}) &= p(\mathbf{x}, \mathbf{z}|\mathbf{y}; \rho^2) \\ &\propto \exp \left[-f(\mathbf{x}, \mathbf{y}) - \beta g(\mathbf{z}) - \frac{1}{2\rho^2} \|\mathbf{x} - \mathbf{z}\|^2 \right] \end{aligned} \quad (13)$$

where ρ is a positive parameter that controls the dissimilarity between \mathbf{x} and \mathbf{z} . This data augmentation (13) is approximate in the sense that the marginal distribution

$$\begin{aligned} \pi_\rho(\mathbf{x}) &= \int_{\mathbb{R}^n} \pi_\rho(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\ &\propto \int_{\mathbb{R}^n} \exp \left[-f(\mathbf{x}, \mathbf{y}) - \beta g(\mathbf{z}) - \frac{1}{2\rho^2} \|\mathbf{x} - \mathbf{z}\|^2 \right] d\mathbf{z} \end{aligned} \quad (14)$$

coincides with the target distribution $\pi(\mathbf{x})$ only in the limiting case $\rho \rightarrow 0$. The conditional distributions¹ associated to the

¹The conditional distributions associated to $\pi_\rho(\mathbf{x}, \mathbf{z})$ are $p(\mathbf{x}|\mathbf{y}, \mathbf{z}; \rho^2)$ and $p(\mathbf{z}|\mathbf{x}; \rho^2)$. To lighten the notations, the coupling parameter ρ^2 will be omitted in what follows, i.e., one employs $p(\mathbf{x}|\mathbf{y}, \mathbf{z})$ and $p(\mathbf{z}|\mathbf{x})$.

augmented posterior $\pi_\rho(\mathbf{x}, \mathbf{z})$ are given by

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) \propto \exp \left[-f(\mathbf{x}, \mathbf{y}) - \frac{1}{2\rho^2} \|\mathbf{x} - \mathbf{z}\|^2 \right] \quad (15)$$

$$p(\mathbf{z}|\mathbf{x}) \propto \exp \left[-\beta g(\mathbf{z}) - \frac{1}{2\rho^2} \|\mathbf{x} - \mathbf{z}\|^2 \right]. \quad (16)$$

The so-called split Gibbs sampler (SGS) alternatively samples according to these two conditional distributions to generate samples asymptotically distributed according to (13) [24], [27]. Interestingly, this splitting allows the two terms $f(\cdot, \mathbf{y})$ and $g(\cdot)$ defining the full potential to be dissociated and involved into two distinct conditional distributions. SGS shares strong similarities with ADMM and HQS methods and is expected to lead to simpler, scalable and more efficient sampling schemes.

Specifically, sampling according to the conditional posterior (15) can be interpreted as solving the initial estimation problem defined by the likelihood function (1) with now a Gaussian distribution with mean \mathbf{z} and diagonal covariance matrix $\rho^2 \mathbf{I}_n$ assigned as a prior. As stated earlier, a large family of imaging inverse problems, such as deblurring, inpainting and super-resolution is characterized by the quadratic potential function $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2\sigma^2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ also considered in this work. This leads to the Gaussian conditional distribution

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}(\mathbf{z}), \mathbf{Q}^{-1}) \quad (17)$$

where the precision matrix \mathbf{Q} and the mean vector $\boldsymbol{\mu}(\cdot)$ are given by

$$\begin{cases} \mathbf{Q} &= \frac{1}{\sigma^2} \mathbf{A}^\top \mathbf{A} + \frac{1}{\rho^2} \mathbf{I} \\ \boldsymbol{\mu}(\mathbf{z}) &= \mathbf{Q}^{-1} \left(\frac{1}{\sigma^2} \mathbf{A}^\top \mathbf{y} + \frac{1}{\rho^2} \mathbf{z} \right). \end{cases} \quad (18)$$

In this case, sampling according to this conditional distribution can be efficiently achieved using dedicated algorithms that depend on the structure of the precision matrix \mathbf{Q} . Interested readers are invited to consult [28] for a recent overview of these methods. It is worth noting that when the potential function $f(\cdot, \mathbf{y})$ is not quadratic, the proposed framework can embed proximal Monte Carlo algorithms to sample from (15), as in [29], [30].

In the specific case considered in this work where $g(\cdot) = g_{\text{red}}(\cdot)$, the conditional distribution (16) can be interpreted as the posterior distribution associated to a Bayesian denoising problem equipped with a RED prior. Its objective boils down to inferring an object \mathbf{z} from the observations \mathbf{x} contaminated by an additive white Gaussian noise with a covariance matrix $\rho^2 \mathbf{I}_n$. Sampling according to this conditional is not straightforward, in particular due to the regularization potential $g_{\text{red}}(\cdot)$ whose definition involves the denoiser $D_\nu(\cdot)$. This work proposes to take advantage of the property (9) by sampling from (16) following a Langevin Monte Carlo (LMC) step, i.e.,

$$\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} + \gamma \nabla \log p(\mathbf{z}^{(t)} | \mathbf{x}) + \sqrt{2\gamma} \boldsymbol{\varepsilon}^{(t)} \quad (19)$$

where $\gamma > 0$ is a fixed step-size and $\{\boldsymbol{\varepsilon}^{(t)}\}_{t \in \mathbb{N}}$ is a sequence of independent and identically distributed n -dimensional standard Gaussian random variables. Given the particular form of the

conditional distribution (16) and the property (9), this recursion writes explicitly as

$$\begin{aligned} \mathbf{z}^{(t+1)} &= \left(1 - \gamma\beta + \frac{1}{\rho^2} \right) \mathbf{z}^{(t)} \\ &\quad - \frac{1}{\rho^2} \mathbf{x}^{(t)} + \gamma\beta D_\nu(\mathbf{z}^{(t)}) + \sqrt{2\gamma} \boldsymbol{\varepsilon}^{(t)}. \end{aligned} \quad (20)$$

The proposed so-called Langevin-within-SGS (LwSGS) instantiated to sample according to the RED posterior (12) is summarized in Algo 1.

Algorithm 1 LwSGS to sample from the RED posterior

Input: denoiser $D_\nu(\cdot)$, regularization parameter β , coupling parameter ρ^2 , step-size γ , number of burn-in iterations N_{bi} , total number of iterations N_{MC}

Initialization: $\mathbf{x}^{(0)}, \mathbf{z}^{(0)}$

1: **for** $t = 0$ to $N_{\text{MC}} - 1$ **do**

 % Sampling the splitting variable according to (20)

2: $\boldsymbol{\varepsilon}^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

3: $\mathbf{z}^{(t+1)} = \left(1 - \gamma\beta + \frac{1}{\rho^2} \right) \mathbf{z}^{(t)} - \frac{1}{\rho^2} \mathbf{x}^{(t)} + \gamma\beta D_\nu(\mathbf{z}^{(t)}) + \sqrt{2\gamma} \boldsymbol{\varepsilon}^{(t)}$

 % Sampling the variable of interest according to (15)

4: $\mathbf{x}^{(t+1)} \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{z}^{(t+1)}), \mathbf{Q}^{-1})$

5: **end for**

Output: collection of samples $\{\mathbf{x}^{(t)}, \mathbf{z}^{(t)}\}_{t=N_{\text{bi}}+1}^{N_{\text{MC}}}$

Because of the discretization followed by the LMC step (19), the samples produced by (20) are biased and are not exactly distributed according to (16). To mitigate this bias and ensure that LMC exactly targets (16), one well-admitted strategy consists in including a Metropolis-Hasting (MH) step, resulting in the Metropolis adjusted Langevin algorithm [31]. Then, combined with the sampling according to (15), the overall resulting sampling algorithm would become a canonical instance of Metropolis-within-Gibbs algorithm whose samples would be ensured to be distributed according to the augmented posterior distribution (13). However, performing this MH step within each iteration of the SGS requires to compute multiple corresponding MH ratios and to accept or reject the proposed samples, which may significantly increase the computational burden of the SGS. In this work, one proposes to bypass this MH correction, yet at the price of an approximation which is controlled. Indeed, the bias induced by the use of a LMC step within a SGS iteration will be investigated in the theoretical analysis conducted in Section III-C.

B. Related Monte Carlo algorithms

The proposed LwSGS algorithm shares some similarities with some recently developed Monte Carlo algorithms, mainly motivated by the will of conducting distributed Bayesian inference over several computer nodes. In [32], the authors adopt a splitting strategy offered by AXDA to derive a distributed SGS (DSGS) when the posterior distribution comprises multiple composite terms. As with LwSGS, the core idea of DSGS can be sketched as replacing the exact sampling of one conditional distribution of the augmented posterior distribution

by a more efficient surrogate sampling technique. Given the particular form of the posterior distributions considered in [32], a suitable choice of this surrogate is shown to be one step of a proximal stochastic gradient Langevin algorithm (PSGLA) [33]. When this sampling is not corrected by a MH step, this leads to an inexact instance of SGS coined as PSGLA-within-SGS. The synchronous distributed version of PSGLA-within-SGS accounts for the hypergraph structure of the involved composite terms to efficiently distribute the variables over multiple workers under controlled communication costs. In the same vein, another synchronous distributed MCMC algorithm referred to as DG-LMC has been introduced in [34] to conduct Bayesian inference when the target log-posterior also writes as a sum of multiple composite terms. Also leveraging AXDA, it adopts a splitting scheme different from LwSGS and PSGLA-within-SGS. Yet, it can be interpreted as an inexact SGS for which multiple conditional distributions are approximately sampled thanks to LMC steps. While the existence of a stationary distribution targeted by PSGLA-within-SGS and its convergence have not been demonstrated in [32], such a thorough theoretical analysis has been conducted for DG-LMC in [34]. Even if these results are a precious asset to conduct a similar analysis of LwSGS, they should be carefully adapted to fit the splitting scheme adopted by LwSGS.

C. Theoretical analysis

This section provides theoretical insights regarding the proposed LwSGS algorithm. For sake of generality and unless otherwise stated, this analysis is conducted for any regularization potential $g(\cdot)$ satisfying assumptions introduced below. However, these assumptions will be also discussed and examined under the prism of the RED paradigm, i.e., with $g(\cdot) = g_{\text{red}}(\cdot)$ and when the proposed LwSGS aims at targeting the RED posterior distribution (12).

As stated above, because of the absence of MH correction after the LMC step, Algo. 1 does not fall into the class of Metropolis-within-Gibbs samplers. Thus, the primer objective of this analysis is to demonstrate that the samples produced by Algo. 1 are asymptotically distributed according to a unique invariant distribution² $\pi_{\rho,\gamma}$ following an ergodic transition kernel denoted $P_{\rho,\gamma}$. Thanks to an appropriate synchronous coupling, the convergence analysis of LwSGS reduces to that of the Markov chain produced by the sampling (20) according to the conditional distribution $p(\mathbf{z}|\mathbf{x})$. One first introduces and discusses two assumptions regarding the regularization potential $g(\cdot)$.

Assumption 2 (Twice differentiability). *The potential function $g(\cdot)$ is twice continuously differentiable and there exists $M_g > 0$ such that $\forall \mathbf{z} \in \mathbb{R}^n$, $\|\nabla^2 g(\mathbf{z})\| \leq M_g$.*

As stated in Section II-A, under the RED conditions (C1) and (C3), the gradient of the RED potential is given by (9). This implies that the regularization potential $g_{\text{red}}(\cdot)$ is twice continuously differentiable with Hessian matrix $\nabla^2 g_{\text{red}}(\mathbf{z}) = \mathbf{I}_n - \nabla D_\nu(\mathbf{z})$. Moreover, thanks to the Jacobian symmetry (C3)

²With a slight abuse of notation, one uses the same notations for a probability distribution and its associated pdf.

and strong passivity (C4) conditions, one has for all $\mathbf{z} \in \mathbb{R}^n$, $\|\nabla^2 g_{\text{red}}(\mathbf{z})\| \leq 2$. In other words, Assumption 2 always holds for RED.

Assumption 3 (Strong convexity). *The potential function $g(\cdot)$ is m_g -strongly convex, i.e., there exists $m_g > 0$ such that $m_g \mathbf{I}_n \preceq \nabla^2 g$.*

In the RED framework, a sufficient condition for the strong convexity of the potential $g_{\text{red}}(\cdot)$ is to ensure that the denoiser $D_\nu(\cdot)$ is contractive, i.e., $\forall (\mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^n \times \mathbb{R}^n$, $\|D_\nu(\mathbf{z}_1) - D_\nu(\mathbf{z}_2)\|_2 \leq \epsilon \|\mathbf{z}_1 - \mathbf{z}_2\|_2$ for some Lipschitz constant $\epsilon < 1$. Under this condition, $g_{\text{red}}(\cdot)$ can be shown to be m_g -strongly convex with $m_g = 1 - \epsilon > 0$. Unfortunately, most existing denoisers do not follow this contraction property [14]. To ensure the strong convexity of the potential $g_{\text{red}}(\cdot)$ when using a deep network-based denoiser, one solution would consist in explicitly including a regularization term into the training loss which constrains the Lipschitz constant [13], [35]. Finally, it is worth noting that when Assumption 3 is satisfied, Assumption 1 is also satisfied, which implies that $p_{\text{red}}(\cdot)$ is well-defined.

Under these assumptions, the convergence of the proposed LwSGS algorithm is stated in the following proposition.

Proposition 2. *Let $\gamma \in \mathbb{R}_+^*$ such that $\gamma \leq (\beta M_g + 1/\rho^2)^{-1}$. Then, under Assumptions 2 and 3, the kernel $P_{\rho,\gamma}$ admits a unique stationary distribution $\pi_{\rho,\gamma}$. Moreover, for any $\mathbf{v} = (\mathbf{x}, \mathbf{z})^\top \in \mathbb{R}^n \times \mathbb{R}^n$ and any $t \in \mathbb{N}^*$, we have*

$$W_2^2(\delta_{\mathbf{v}} P_{\rho,\gamma}^t, \pi_{\rho,\gamma}) \leq C_1 (1 - \gamma \beta m_g)^{2(t-1)} W_2^2(\delta_{\mathbf{v}}, \pi_{\rho,\gamma}),$$

where $C_1 = 1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2$.

Proof. See Appendix B. □

The rate of convergence of the proposed sampler is given by $1 - \gamma \beta m_g$. The asymptotic convergence of the samples produced by the kernel $P_{\rho,\gamma}$ towards the distribution $\pi_{\rho,\gamma}$ is only possible if the LMC discretization step γ is sufficiently small. The formula above establishes convergence for any step-size $\gamma \leq (\beta M_g + 1/\rho^2)^{-1}$. From Assumptions 2 and 3, $-\log p(\mathbf{z}|\mathbf{x})$ is βm_g -strongly convex and $(\beta M_g + 1/\rho^2)$ -smooth, i.e., $\beta m_g \mathbf{I}_n \preceq -\nabla^2 \log p(\mathbf{z}|\mathbf{x}) \preceq (\beta M_g + 1/\rho^2) \mathbf{I}_n$. From [34] and [26], a sufficient condition on the step size γ to ensure contraction w.r.t. the Wasserstein distance is $\gamma \leq 2/(\beta m_g + \beta M_g + 1/\rho^2)$. Thus the coupling parameter ρ implicitly determines the convergence rate. The smaller ρ , the smaller γ and the slower the convergence is.

Once the asymptotic convergence of the samples produced by Algo. 1 has been ensured, the second stage of the theoretical analysis consists in analyzing the bias between the stationary distribution $\pi_{\rho,\gamma}$ and the targeted augmented distribution π_ρ . This bias, which results from the use of the LMC step in Algo. 1 to target the conditional distribution $p(\mathbf{z}|\mathbf{x})$ defined by (16), is quantified in the following proposition.

Proposition 3. *Let $\gamma \in \mathbb{R}_+^*$ such that $\gamma \leq 2(\beta m_g + \beta M_g + 1/\rho^2)^{-1}$. Then, under Assumptions 2 and 2, we have*

$$W_2^2(\pi_\rho, \pi_{\rho,\gamma}) \leq n \gamma C_2 \tilde{M}^2 \left(1 + \frac{\gamma^2 \tilde{M}^2}{12} + \frac{\gamma \tilde{M}^2}{2\tilde{m}} \right),$$

where $\tilde{m} = \beta m_g + 1/\rho^2$, $\tilde{M} = \beta M_g + 1/\rho^2$ and $C_2 = \frac{2}{\beta m_g} (1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2)$.

Proof. See Appendix C. \square

The bias is upper-bounded by a term driven by the step size. More precisely, as γ decreases towards zero, the squared Wasserstein distance $W_2^2(\pi_\rho, \pi_{\rho,\gamma})$ is upper-bounded by $\mathcal{O}(n\gamma)$. Thus, for a sufficiently small step size γ , LwSGS produces a Markov chain with a stationary distribution $\pi_{\rho,\gamma}$ that can be arbitrary close to π_ρ . This result is in agreement with those stated in [34, Proposition 4] and [26, Corollary 7].

D. Revisiting PnP-ULA and AXDA from the RED paradigm

This section draws connections between AXDA, the proposed algorithm and PnP-ULA [22]. As a reminder, in a nutshell, PnP-ULA targets a posterior distribution of the form

$$p_\epsilon(\mathbf{x}|\mathbf{y}) \propto \exp[-f(\mathbf{x}, \mathbf{y})] p_\epsilon(\mathbf{x}) \quad (21)$$

with $\epsilon > 0$ where

$$p_\epsilon(\mathbf{x}) \propto \int_{\mathbb{R}^n} p(\mathbf{z}) \exp\left[-\frac{1}{2\epsilon} \|\mathbf{x} - \mathbf{z}\|^2\right] d\mathbf{z} \quad (22)$$

defines the regularized counterpart of the prior distribution $p(\mathbf{x})$. Interestingly, when the prior writes $p(\mathbf{x}) \propto \exp[-\beta g(\mathbf{x})]$ and $\rho^2 = \epsilon$, the posterior distribution $p_\epsilon(\mathbf{x}|\mathbf{y})$ targeted by PnP-ULA (21) perfectly matches the marginal distribution $\pi_\rho(\mathbf{x})$ in (14) resulting from an AXDA strategy and targeted by a SGS.

Besides, from an algorithmic point of view, ULA recursions applied to (21) conventionally write

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma \nabla f(\mathbf{x}^{(t)}, \mathbf{y}) + \gamma \nabla \log p_\epsilon(\mathbf{x}^{(t)}) + \sqrt{2\gamma} \boldsymbol{\varepsilon}^{(t)}.$$

Thanks to the Tweedie's identity [25], the score function can be replaced by the denoising residual, i.e., $\epsilon \nabla \log p_\epsilon(\mathbf{x}) = \mathbf{D}_\epsilon^*(\mathbf{x}) - \mathbf{x}$ where $\mathbf{D}_\epsilon^*(\cdot)$ is an MMSE denoiser. This leads to the (simplified) PnP-ULA scheme

$$\begin{aligned} \text{(PnP-ULA)} : \quad \mathbf{x}^{(t+1)} &= \mathbf{x}^{(t)} - \gamma \nabla f(\mathbf{x}^{(t)}, \mathbf{y}) \\ &+ \frac{\gamma}{\epsilon} \left(\mathbf{D}_\epsilon^*(\mathbf{x}^{(t)}) - \mathbf{x}^{(t)} \right) + \sqrt{2\gamma} \boldsymbol{\varepsilon}^{(t)}. \end{aligned} \quad (23)$$

This simplified scheme departs from the canonical PnP-ULA scheme studied in [22] by only omitting an additional term $\frac{\gamma}{\lambda} [\Pi_{\mathbb{S}}(\mathbf{x}^{(t)}) - \mathbf{x}^{(t)}]$ where $\Pi_{\mathbb{S}}(\cdot)$ denotes the projection onto the convex and compact set \mathbb{S} . This term has been included into the PnP-ULA scheme for technical reasons to derive convergence results. Its impact will be empirically shown to be marginal in practice (see Section IV-F). Conversely, using (9), ULA recursions applied to the RED posterior (12) writes

$$\begin{aligned} \text{(RED-ULA)} : \quad \mathbf{x}^{(t+1)} &= \mathbf{x}^{(t)} - \gamma \nabla f(\mathbf{x}^{(t)}, \mathbf{y}) \\ &+ \gamma \beta \left(\mathbf{D}_\nu(\mathbf{x}^{(t)}) - \mathbf{x}^{(t)} \right) + \sqrt{2\gamma} \boldsymbol{\varepsilon}^{(t)}. \end{aligned} \quad (24)$$

It clearly appears that the RED-ULA scheme defined by the previous recursion (24) coincides with PnP-ULA (23) when $\beta = \frac{1}{\epsilon}$ and the denoiser embedded into RED is chosen as the MMSE denoiser, i.e., $\mathbf{D}_\nu(\cdot) = \mathbf{D}_\epsilon^*(\cdot)$. Since in practice the use of an MMSE denoiser is infeasible, PnP-ULA is implemented with an off-the-shelf denoiser. Thus, RED-ULA is no more than the practical implementation of PnP-ULA.

Moreover, an important corollary deals with the respective prior distributions defining the posteriors targeted by the three considered algorithms, namely PnP-ULA, RED-ULA and RED-SGS³. Indeed, on one hand, PnP-ULA and RED-ULA target the same posterior distribution (21) (provided the use of an MMSE denoiser). On another hand, PnP-ULA and RED-SGS also target the same posterior distribution (12). It yields that the regularized prior implicitly induced by AXDA coincides with the RED prior based on an MMSE denoiser, i.e., $p_\epsilon(\mathbf{x}) = p_{\text{red}}^*(\mathbf{x})$ and, in particular,

$$-\log p_\epsilon(\mathbf{x}) = \frac{1}{2\epsilon} \mathbf{x}^\top [\mathbf{x} - \mathbf{D}_\epsilon^*(\mathbf{x})]. \quad (25)$$

Note that differentiating this identity obviously leads again to the celebrated Tweedie's identity.

IV. EXPERIMENTS

A. Experimental setup

Experiments have been conducted based on two popular image data sets, namely the Flickr Faces High Quality (FFHQ) data set [36] and the ImageNet data set [37]. All images are RGB images of size 256×256 pixels ($n = 256^2$) and have been normalized to the range $[0, 1]$. The performance of the proposed RED-LwSGS algorithm is assessed w.r.t. to three inversion tasks

- *deblurring*: the operator \mathbf{A} is assumed to be a $n \times n$ circulant convolution matrix associated with a spatially invariant blurring kernel. It is chosen as a Gaussian kernel of size 25×25 with standard deviation 1.6.
- *inpainting*: the operator \mathbf{A} stands for a binary mask with $m \ll n$. It is designed such that 80% of the total pixels are randomly masked across the three color channels.
- *super-resolution*: the operator \mathbf{A} is decomposed as $\mathbf{A} = \mathbf{S}\mathbf{B}$ where the $n \times n$ matrix \mathbf{B} stands for a spatially invariant Gaussian blur of size 7×7 with standard deviation 1.6 and the operator \mathbf{S} is a $m \times n$ binary matrix which performs a regular subsampling of factor $d = 4$ in each dimension (i.e., $m = nd^2$). It is worth noting that for this task, the AXDA model needs to be adapted to account for the specific structure of the matrix \mathbf{A} (more details are given in Appendix D).

For all tasks, the degraded images have been corrupted by an additive Gaussian noise to reach a signal-to-noise ratio of SNR = 30dB.

B. Compared methods

The proposed RED-LwSGS algorithm has been implemented using the pre-trained deep network DRUNet [11] as the denoiser $\mathbf{D}_\nu(\cdot)$. It has been taken directly from the corresponding repository and has been used without further fine-tuning for the considered inversion tasks, see Appendix E-A for complementary information. The test images have never been seen by the model while training to avoid any bias due to

³The algorithmic scheme denoted RED-SGS is defined as a canonical Gibbs sampler with targets the augmented distribution (13). It can be interpreted as the RED-LwSGS for which the sampling according to (16) would be exact.

potentially overfitted pre-trained models. The proposed method has been compared to several state-of-the-art inversion methods:

- RED-ADMM [15]: ADMM with RED regularization based on the same DRUNet denoiser;
- RED-HQS: HQS algorithm with RED regularization based on the same DRUNet denoiser;
- PnP-ADMM [5]: ADMM with a PnP regularization based on the same DRUNet denoiser;
- PnP-ULA [22]: ULA with a PnP regularization based on the same DRUNet denoiser;
- TV-SP [24]: SGS with a TV regularization;
- TV-MYULA [38]: Moreau-Yosida ULA with TV regularization;
- DiffPIR [39]: denoising diffusion model for PnP image restoration.

It is worth noting that RED-ADMM, RED-HQ and PnP-ADMM are optimization-based methods which provide only point estimates of the restored images. Conversely, PnP-ULA, TV-SP and TV-MYULA are Monte Carlo methods that are able to enrich point estimates with credibility intervals. For these sampling methods, the results reported below correspond to the MMSE estimates approximated by averaging the generated samples after the burn-in period. Complementary information regarding the algorithm implementations are reported in Appendices E-B and E-C.

C. Figures-of-merit

Beyond performing visual inspection, the methods are compared with respect to several quantitative figures-of-merit. Peak signal noise ratio (PSNR) (dB) and structural similarity index (SSIM) are considered as image quality metrics (the higher the score, the better the reconstruction). They are complemented with a perceptual metric, namely the learned perceptual image patch similarity (LPIPS), for which the lower the score, the better the reconstruction. Moreover, to assess the effectiveness of the sampling algorithms studied, they are also compared in term of integrated autocorrelation time (IAT), which is also an indicator for good or bad mixing (the lower, the better) [40]. Finally, all methods are compared in terms of computational times when the algorithms are implemented on a server equipped with 48 CPU cores Intel 2.8Ghz, 384Go RAM, Nvidia A100 GPU.

D. Experimental results

The results obtained by the compared algorithms when performing the three considered tasks are reported in Tables I and II for the two datasets FFHQ and ImageNet, respectively. These results show that the proposed RED-LwSGS method achieves very competitive performance for all three tasks. In the case of inpainting, the forward operator masking 80% pixels is non-invertible and the problem is expected to require further prior regularization than the two other tasks. For this task, algorithms relying on data-driven regularizations, such as RED-LwSGS and PnP-ULA, appear to include more informative priors when compared to TV-MYULA and TV-SP which rely on the same model-based regularization. For super-resolution,

which is a more challenging problem than deblurring and inpainting, RED-LwSGS performs similarly to DiffPIR, RED-ADMM and RED-HQS when other MCMC algorithms fail.

Figures 1 and 2 visually assess the performance by depicting the results obtained on test images drawn from the FFHQ and ImageNet datasets. For all the considered tasks, the proposed RED-LwSGS method produces high-quality, realistic images that closely match ground-truth details. As mentioned above, RED-LwSGS generates samples asymptotically distributed according to the posterior distribution. These samples can be used to quantify estimation uncertainty. The rightmost panels in Fig. 1 and 2 illustrate this advantage by depicting the estimated pixelwise standard deviations obtained by the proposed algorithms. It is worth noting that this added value cannot be provided by optimization-based methods such as DiffPIR, RED-ADMM, RED-HQS and PnP-ADMM, which only offer point estimates. Noticeably, pixels located in homogeneous regions are characterized by lower uncertainty, while pixels in textured regions, edges, or complex structures appear to be estimated with more difficulty.

Tables I and II also reported the computational times required by the compared algorithms when tackling each restoration tasks. Among the class of sampling methods, RED-LwSGS stands out for its smallest computational times. Moreover, as for the other sampling-based methods, RED-LwSGS remains within a factor of 50 compared to the optimization-based methods, namely DiffPIR, RED-ADMM, RED-HQS, and PnP-ADMM. The price to pay for offering an uncertainty quantification on top of point estimation seems reasonable.

Finally, the convergence properties of the compared sampling-based algorithms have been assessed by monitoring the autocorrelation function (ACF) of the median components of the chains generated by those algorithms. By denoting $\mathbf{x}^{(t)} = [x_1^{(t)}, \dots, x_n^{(t)}]^\top$, the median component has been defined as the produced pixelwise Monte Carlo chain $\{x_i^{(t)}\}_{t=N_{\text{bi}}+1}^{N_{\text{MC}}}$ with the median variance. Faster decreasing ACF means that the samples are less correlated and generally implies faster convergence of the Markov chain. Fig. 3 depicts these ACFs for the three restoration tasks conducted on one image from the FFHQ data set. For the deblurring task, it is not clear which of the compared methods is the more efficient, i.e., with the fastest ACF decay. Conversely, for the inpainting and super-resolution tasks, the ACF of RED-LwSGS decreases faster than the ACFs obtained with the two other Monte Carlo algorithms. This finding is confirmed by the IAT measures reported in Tables I and II.

E. Does DRUNet meet the RED conditions?

The explicit formula (9) of the RED gradient requires the RED conditions (C1)–(C4) to be satisfied, i.e., the denoiser $D_\nu(\cdot)$ should be differentiable, locally homogeneous, with symmetric Jacobian and strongly passive. It is legitimate to assess whether these conditions are verified for the deep denoiser considered in the experiments, namely DRUNet. First, the canonical implementation of DRUNet makes $D_\nu(\cdot)$ not continuously differentiable w.r.t. the input due to the use of

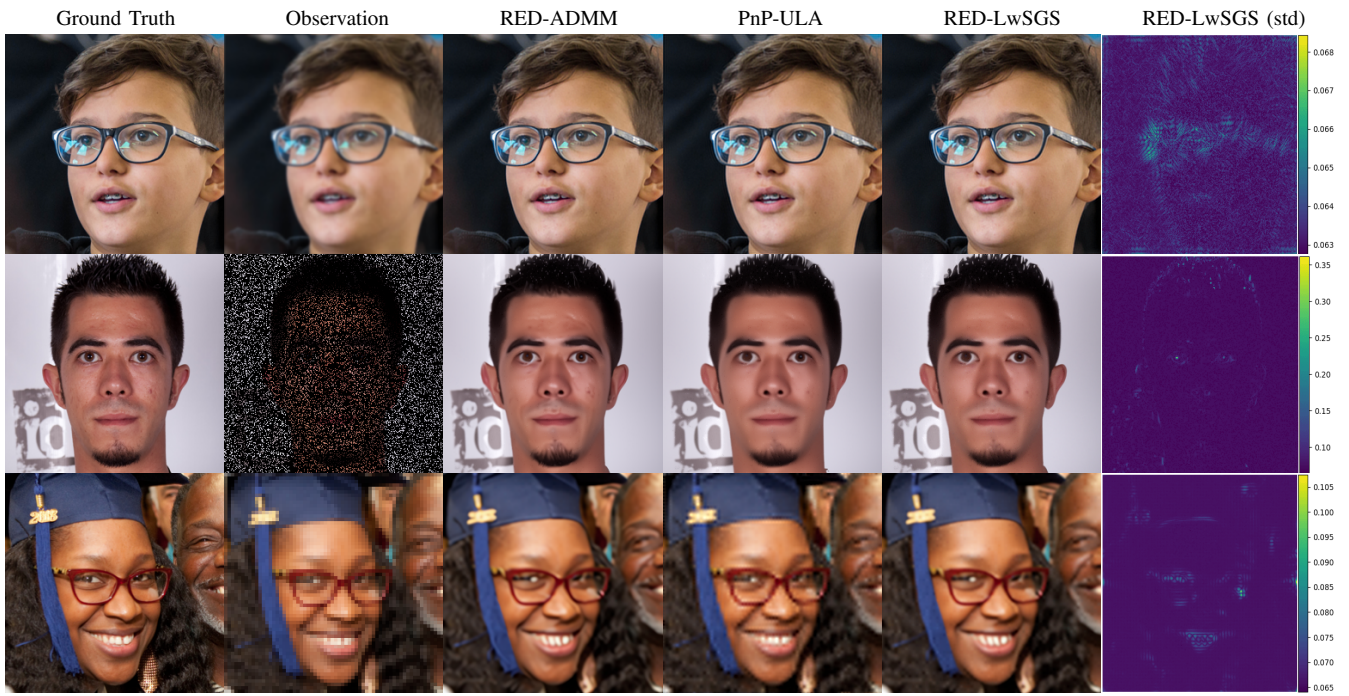


Fig. 1. FFHQ data set: images recovered by the compared methods for deblurring (top), inpainting (middle) and super-resolution (bottom).

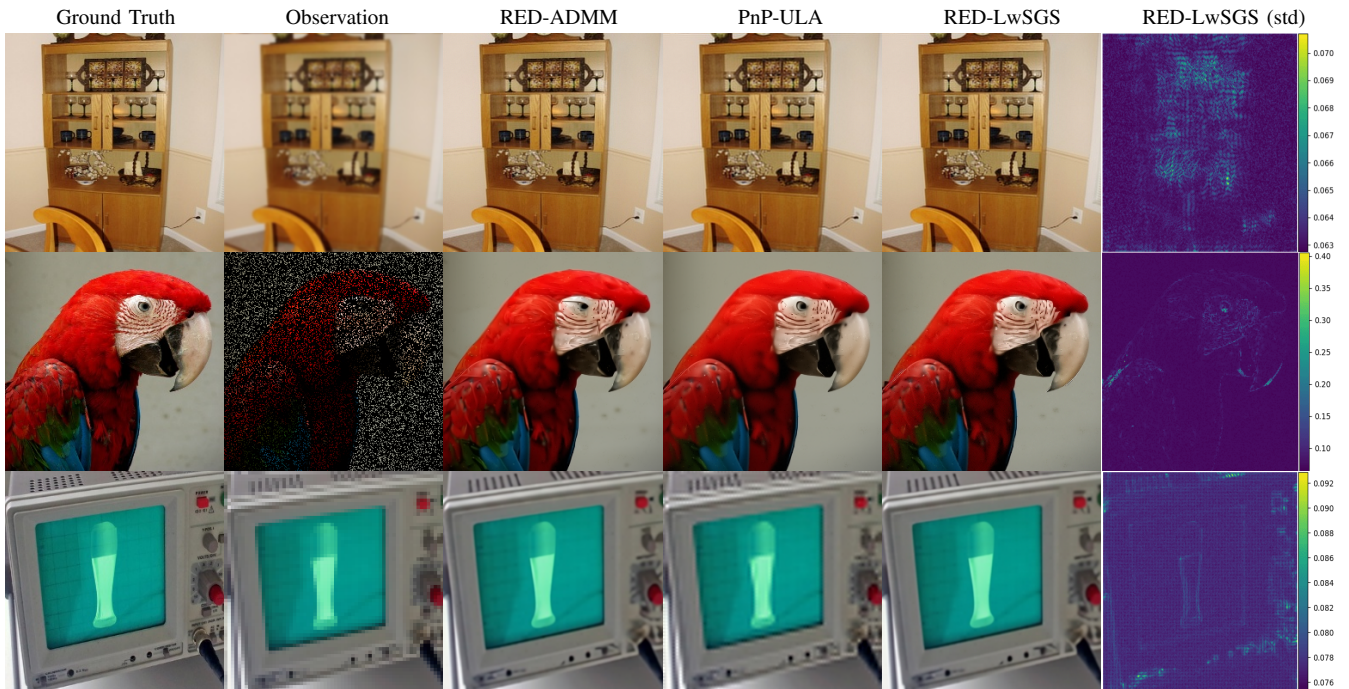


Fig. 2. ImageNet data set: images recovered by the compared methods for deblurring (top), inpainting (middle) and super-resolution (bottom).

TABLE I

FFHQ DATA SET: AVERAGE PERFORMANCE OVER A TEST SET OF 100 IMAGES AND CORRESPONDING STANDARD DEVIATIONS. **BOLD**: BEST SCORE, UNDERLINE: SECOND SCORE.

	Observation	RED-LwSGS	PnP-ULA	TV-MYULA	TV-SP	RED-ADMM	RED-HQS	PnP-ADMM	DiffPIR	
Deblurring	PSNR(dB) ↑	29.8550	40.13±1.931	38.06±1.986	38.03±2.028	39.89±1.320	41.28±2.459	40.80±2.560	34.03±7.601	37.15±2.066
	SSIM ↑	0.8994	0.982±0.003	0.975±0.006	0.981±0.006	0.970±0.004	0.985±0.004	0.984±0.005	0.869±0.154	0.957±0.010
	LPIPS ↓	0.0487	0.003±0.002	0.005±0.004	0.005±0.004	<u>0.002±0.001</u>	0.002±0.002	0.002±0.002	0.033±0.080	0.007±0.001
	IAT ↓	-	<u>75.17±12.42</u>	75.21±12.43	75.22±12.42	75.14±12.42	-	-	-	-
	Time(s) ↓	-	44±5	52±7	64±5	40±5	2±0	2±0	2±0	1±0
Inpainting	PSNR(dB) ↑	7.2069	30.73±2.932	<u>31.46±2.650</u>	27.71±1.881	27.27±1.781	31.63±2.672	31.36±2.293	31.32±3.142	31.26±2.25
	SSIM ↑	0.0678	0.908±0.028	0.906±0.026	0.830±0.040	0.815±0.041	0.911±0.024	0.901±0.025	0.915±0.042	0.890±0.025
	LPIPS ↓	0.5831	0.023±0.017	<u>0.020±0.013</u>	0.056±0.026	0.061±0.027	0.019±0.013	0.021±0.014	0.019±0.015	0.021±0.005
	IAT ↓	-	<u>75.24±12.48</u>	<u>75.51±12.55</u>	76.97±13.51	113.86±12.17	-	-	-	-
	Time(s) ↓	-	74±1	79±1	150±7	71±1	3±0	2±0	2±0	2±0
Super-res.	PSNR(dB) ↑	-	30.43±2.161	29.01±2.013	28.99±2.017	28.94±2.019	30.49±2.222	<u>30.54±2.206</u>	30.13±2.184	30.99±2.212
	SSIM ↑	-	0.872±0.036	0.847±0.037	0.847±0.037	0.846±0.037	<u>0.875±0.036</u>	0.876±0.036	0.867±0.037	0.868±0.034
	LPIPS ↓	-	0.035±0.021	0.050±0.024	0.049±0.024	0.051±0.024	<u>0.034±0.020</u>	<u>0.034±0.020</u>	0.035±0.021	0.011±0.008
	IAT ↓	-	75.75±12.62	75.88±12.64	<u>75.85±12.64</u>	75.86±12.63	-	-	-	-
	Time(s) ↓	-	115±25	128±40	133±26	112±23	3±1	3±1	3±1	2±0

TABLE II

IMAGENET DATA SET: AVERAGE PERFORMANCE OVER A TEST SET OF 100 IMAGES AND CORRESPONDING STANDARD DEVIATIONS. **BOLD**: BEST SCORE, UNDERLINE: SECOND SCORE.

	Observation	RED-LwSGS	PnP-ULA	TV-MYULA	TV-SP	RED-ADMM	RED-HQS	PnP-ADMM	DiffPIR	
Deblurring	PSNR(dB) ↑	26.6362	<u>34.85±3.961</u>	32.39±4.063	31.89±4.075	34.58±4.026	35.41±4.024	34.48±3.062	31.63±7.783	31.89±4.139
	SSIM ↑	0.7676	<u>0.955±0.024</u>	0.925±0.046	0.921±0.051	0.954±0.026	0.956±0.024	0.914±0.064	0.867±0.171	0.896±0.065
	LPIPS ↓	0.1312	0.018±0.019	0.036±0.037	0.041±0.040	0.020±0.021	<u>0.016±0.019</u>	0.012±0.011	0.059±0.108	0.044±0.016
	IAT ↓	-	66.29±19.98	66.76±19.89	66.89±19.89	<u>66.34±19.98</u>	-	-	-	-
	Time(s) ↓	-	44±3	50±2	69±32	43±1	2±1	2±1	2±1	1±0
Inpainting	PSNR(dB) ↑	7.3659	25.74±4.461	26.45±4.208	24.82±3.431	24.57±3.301	<u>26.52±4.390</u>	26.29±4.377	26.27±4.387	26.87±3.97
	SSIM ↑	0.0632	0.774±0.136	0.777±0.125	0.697±0.129	0.684±0.128	0.787±0.119	0.779±0.120	0.769±0.132	0.765±0.123
	LPIPS ↓	0.5824	0.097±0.082	0.088±0.078	0.136±0.084	0.144±0.084	<u>0.085±0.074</u>	0.090±0.081	0.089±0.079	0.035±0.034
	IAT ↓	-	66.53±19.79	<u>67.99±19.52</u>	113.60±18.53	107.55±18.57	-	-	-	-
	Time(s) ↓	-	75±2	80±3	163±14	73±1	3±0	3±0	3±0	8±0
Super-res.	PSNR(dB) ↑	-	<u>26.24±3.933</u>	25.42±3.461	25.35±3.445	25.39±3.445	26.22±3.936	26.21±3.926	25.84±3.899	26.31±3.791
	SSIM ↑	-	0.722±0.148	0.693±0.142	0.687±0.146	0.693±0.141	<u>0.721±0.148</u>	<u>0.721±0.148</u>	0.707±0.150	0.712±0.142
	LPIPS ↓	-	<u>0.120±0.089</u>	0.142±0.090	0.148±0.092	0.142±0.089	<u>0.120±0.090</u>	<u>0.120±0.089</u>	0.123±0.090	0.106±0.051
	IAT ↓	-	93.36±27.81	98.38±29.16	102.11±28.45	<u>97.25±28.12</u>	-	-	-	-
	Time(s) ↓	-	112±22	119±30	129±15	109±13	3±0	3±0	3±0	3±0

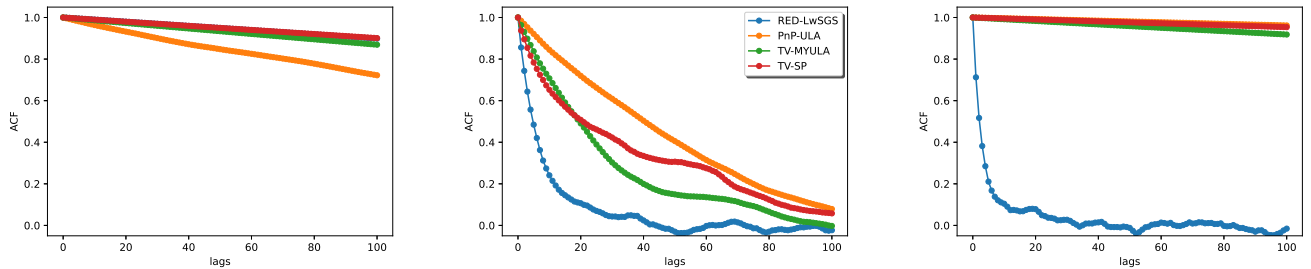


Fig. 3. FFHQ data set: (absolute) autocorrelation function (ACF) of the samples generated by the compared algorithms for deblurring (left), inpainting (middle) and super-resolution (right).

TABLE III
NUMERICAL EXPERIMENTS TO ASSESS THE RED CONDITIONS.

		FFHQ	ImageNet
Homogeneity	NMSE _{LH,1}	1.78×10^{-8}	6.23×10^{-8}
	NMSE _{LH,2}	1.75×10^{-2}	6.21×10^{-2}
Jacobian symmetry	NMSE _{JS}	5.81×10^{-2}	7.56×10^{-2}
Strong passivity	MSR	1.44	1.76

ReLU activation functions. However, to ensure (C2), the overall architecture can be made continuously differentiable by replacing them with SoftPlus activation functions, which are C^∞ , as suggested in [14]. Besides, the three other RED conditions are empirically assessed through numerical experiments. The local homogeneity condition (C1) of the denoiser is evaluated by computing the two following normalized mean square errors [16]

$$\text{NMSE}_{\text{LH},1} = \mathbb{E} \left[\frac{\|D_\nu((1+\epsilon)\mathbf{x}) - (1+\epsilon)D_\nu(\mathbf{x})\|^2}{\|(1+\epsilon)D_\nu(\mathbf{x})\|^2} \right]$$

and

$$\text{NMSE}_{\text{LH},2} = \mathbb{E} \left[\frac{\|\nabla D_\nu(\mathbf{x})\mathbf{x} - D_\nu(\mathbf{x})\|^2}{\|D_\nu(\mathbf{x})\|^2} \right]$$

which are motivated by the definition (6) and the property (7), respectively. These metrics should be close to zero to ensure local homogeneity. Further, following (C3), the denoiser should have a symmetric Jacobian. This characteristic is empirically evaluated by computing

$$\text{NMSE}_{\text{JS}} = \mathbb{E} \left[\frac{\|\nabla D_\nu(\mathbf{x}) - \nabla D_\nu^\top(\mathbf{x})\|^2}{\|\nabla D_\nu(\mathbf{x})\|^2} \right]$$

which should be close to zero for a symmetric Jacobian [16]. Finally, to assess the strong passivity condition (C4), one considers the mean spectral radius

$$\text{MSR} = \mathbb{E} [\eta(\nabla D_\nu(\mathbf{x}))]$$

which can be computed by the power iteration and should be smaller than 1. When computing these metrics, the (i, j) -element of the gradient of $D_\nu(\cdot)$ has been approximated as

$$[\nabla D_\nu(\mathbf{x})]_{i,j} \approx \frac{[D_\nu(\mathbf{x} + \epsilon \mathbf{e}_j)]_i - [D_\nu(\mathbf{x} - \epsilon \mathbf{e}_j)]_i}{2\epsilon}$$

where \mathbf{e}_j denotes the j th canonical basis vector, i.e., the j th column of \mathbf{I}_n , and $\epsilon > 0$ is small enough. These metrics are reported in Table III when the four considered scores have been computed over 100 patches of size 32×32 extracted from images of the two data sets, namely FFHQ and ImageNet. They show that the DRUNet denoiser seems to satisfy the local homogeneity and symmetric Jacobian conditions. However, DRUNet has a spectral radius greater than 1. Yet, to ensure a strongly passive deep neural network-based denoiser, various strategies could be envisioned, such as spectral normalization [13].

TABLE IV
NUMERICAL COMPARAISONS OF RED-ULA AND PnP-ULA.

		PSNR	SSIM	LPIPS	IAT	% Proj.
RED-ULA		38.03	0.9809	0.0053	75.22	-
	$\mathbb{S} = [0, 1]$	38.46	0.9813	0.0048	75.21	96.71
PnP-ULA	$\mathbb{S} = [-1, 2]$	38.03	0.9809	0.0053	75.22	0

F. RED-ULA vs. PnP-ULA

Section III-D has drawn some connections between PnP-ULA and RED-ULA. In particular, it has shown that in practice the two algorithms basically reduce to the same algorithmic scheme, except that PnP-ULA embeds an additional projection step onto an arbitrary pre-defined set \mathbb{S} . This projection aims at ensuring that the drift satisfies an asymptotic growth condition. To experimentally validate this equivalence, Table IV reports the performance of PnP-ULA and RED-ULA when tackling the deblurring task. These performances have been computed over 100 images of the FFHQ dataset. Two configurations for the set \mathbb{S} are considered, $\mathbb{S} = [0, 1]$ and $\mathbb{S} = [-1, 2]$. The rates of activation of the constraint, i.e., the proportion of samples generated by PnP-ULA that do not satisfy the drift condition and should be projected onto the set \mathbb{S} , are also reported in terms of percentage. These results show that, when $\mathbb{S} = [-1, 2]$, the projection embedded in PnP-ULA is never activated and the performance is the same as the one obtained by RED-ULA, which confirms that the two algorithms are identical. When $\mathbb{S} = [0, 1]$, this projection is applied to almost all the samples generated by PnP-ULA, without significantly affecting the performance.

V. CONCLUSION

This work built a Bayesian counterpart of the regularization-by-denoising (RED) engine, offering a data-driven framework to define prior distributions in Bayesian inversion tasks. It defined a new probability distribution from the RED potential, which was subsequently embedded into a Bayesian model as a prior distribution. Since the resulting RED posterior distribution was not standard, a dedicated Monte Carlo algorithm was designed. By leveraging an asymptotically exact data augmentation (AXDA), this algorithm was a particular instance of the split Gibbs sampler which had the great advantage of decoupling the data-fitting term and the RED potential. One stage of SGS was performed following a Langevin Monte Carlo step, which leads to the so-called Langevin-within-split Gibbs sampling. A thorough theoretical analysis was conducted to assess the convergence guarantees of the algorithm. Some tight connections were drawn between AXDA and RED to show that the implicit prior resulting from an AXDA scheme coincides with the RED prior defined by a MMSE denoiser. Extensive numerical experiments showed that the proposed approach competes favorably with state-of-the-art variational and Monte Carlo methods when tackling conventional inversion tasks, namely deblurring, inpainting and super-resolution. The proposed approach was shown to provide a comprehensive characterization of the solutions which could be accompanied

by an uncertainty quantification. By bridging the gap between the RED paradigm and Bayesian inference, this work opened new avenues for incorporating data-driven regularizations into Monte Carlo algorithms.

APPENDIX A
PROOF OF PROPOSITION 1

Under the RED conditions, the RED prior (10) can be rewritten according to the pseudo-quadratic form

$$p_{\text{red}}(\mathbf{x}) \propto \exp \left[-\frac{\beta}{2} \mathbf{x}^\top \mathbf{\Lambda}(\mathbf{x}) \mathbf{x} \right]$$

with $\mathbf{\Lambda}(\mathbf{x}) = \mathbf{I}_n - \nabla D_\nu(\mathbf{x})$. From Assumption 1, there exists $\lambda_{\min} > 0$ such that $\lambda_{\min} \mathbf{I}_n \preceq \mathbf{\Lambda}(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^n$. This implies that $\lambda_{\min} \mathbf{x}^\top \mathbf{x} \leq \mathbf{x}^\top \mathbf{\Lambda}(\mathbf{x}) \mathbf{x}$ and

$$\int_{\mathbb{R}^n} p_{\text{red}}(\mathbf{x}) d\mathbf{x} \leq \int_{\mathbb{R}^n} \exp \left[-\frac{\beta}{2} \lambda_{\min} \|\mathbf{x}\|^2 \right] d\mathbf{x} < \infty.$$

APPENDIX B
PROOF OF PROPOSITION 2

The proposed algorithm LwSGS does not fall into the class of Metropolis-within-Gibbs samplers. Hence, the first step of the analysis consists in demonstrating that the Markov kernel $P_{\rho, \gamma}$ associated with the homogeneous Markov chain produced by LwSGS possesses a unique invariant distribution and is geometrically ergodic. This is achieved through an appropriate synchronous coupling, reducing the convergence analysis of $(\mathbf{V}^{(t)})_{t \in \mathbb{N}} = (\mathbf{X}^{(t)}, \mathbf{Z}^{(t)})_{t \in \mathbb{N}}$ to that of the marginal process $(\mathbf{Z}^{(t)})_{t \in \mathbb{N}}$.

The Markov chain $(\mathbf{V}^{(t)})_{t \in \mathbb{N}}$ is associated with the Markov kernel defined, for any $\mathbf{v} = (\mathbf{x}, \mathbf{z}) \in \mathbb{R}^n \times \mathbb{R}^n$ and $\mathbf{A} \in \mathcal{B}(\mathbb{R}^n)$, $\mathbf{B} \in \mathcal{B}(\mathbb{R}^n)$, by

$$P_{\rho, \gamma}(\mathbf{v}, \mathbf{A} \times \mathbf{B}) = \int_{\mathbf{B}} Q_{\rho, \gamma}(\mathbf{z}, d\tilde{\mathbf{z}}|\mathbf{x}) \int_{\mathbf{A}} p(d\mathbf{x}|\mathbf{y}, \tilde{\mathbf{z}}), \quad (26)$$

where $p(\cdot|\mathbf{y}, \tilde{\mathbf{z}}) \propto \exp \left[-f(\cdot, \mathbf{y}) - \frac{1}{2\rho^2} \|\cdot - \tilde{\mathbf{z}}\|^2 \right]$ and $Q_{\rho, \gamma}(\mathbf{z}, \mathbf{B}|\mathbf{x})$ is the conditional Markov transition kernel given by

$$Q_{\rho, \gamma}(\mathbf{z}, \mathbf{B}|\mathbf{x}) = \int_{\mathbf{B}} \exp \left[-\frac{1}{4\gamma} \left\| \tilde{\mathbf{z}} - b_{\rho, \gamma}(\mathbf{x}, \mathbf{z}) \right\|^2 \right] \frac{d\tilde{\mathbf{z}}}{(4\pi\gamma)^{n/2}}$$

with $b_{\rho, \gamma}(\mathbf{x}, \mathbf{z}) = \left(1 - \frac{\gamma}{\rho^2}\right) \mathbf{z} - \frac{\gamma}{\rho^2} \mathbf{x} + \gamma \nabla g(\mathbf{z})$.

Consider two independent sequences $(\boldsymbol{\xi}^{(t)})_{t \geq 1}$ and $(\boldsymbol{\eta}^{(t)})_{t \geq 1}$ of independent and identically distributed (i.i.d.) n -dimensional standard Gaussian random variables. Let introduce the stochastic processes $(\mathbf{V}^{(t)}, \tilde{\mathbf{V}}^{(t)})_{t \geq 0}$ starting from $(\mathbf{v}^{(0)}, \tilde{\mathbf{v}}^{(0)}) = ((\mathbf{x}, \mathbf{z})^\top, (\tilde{\mathbf{x}}, \tilde{\mathbf{z}})^\top)$ and recursively defined as ($t \geq 0$)

$$\mathbf{V}^{(t+1)} = \left(\mathbf{X}^{(t+1)}, \mathbf{Z}^{(t+1)} \right)^\top \quad (27)$$

$$\tilde{\mathbf{V}}^{(t+1)} = \left(\tilde{\mathbf{X}}^{(t+1)}, \tilde{\mathbf{Z}}^{(t+1)} \right)^\top \quad (28)$$

with

$$\begin{aligned} \mathbf{Z}^{(t+1)} &= \left(1 - \frac{\gamma}{\rho^2}\right) \mathbf{Z}^{(t)} + \frac{\gamma}{\rho^2} \mathbf{X}^{(t)} - \gamma \beta \nabla g(\mathbf{Z}^{(t)}) + \sqrt{2\gamma} \boldsymbol{\eta}^{(t+1)} \\ \tilde{\mathbf{Z}}^{(t+1)} &= \left(1 - \frac{\gamma}{\rho^2}\right) \tilde{\mathbf{Z}}^{(t)} + \frac{\gamma}{\rho^2} \tilde{\mathbf{X}}^{(t)} - \gamma \beta \nabla g(\tilde{\mathbf{Z}}^{(t)}) + \sqrt{2\gamma} \boldsymbol{\eta}^{(t+1)} \end{aligned} \quad (29)$$

and

$$\begin{aligned} \mathbf{X}^{(t+1)} &= \boldsymbol{\mu}(\mathbf{Z}^{(t+1)}) + \mathbf{Q}^{-\frac{1}{2}} \boldsymbol{\xi}^{(t+1)}, \\ \tilde{\mathbf{X}}^{(t+1)} &= \boldsymbol{\mu}(\tilde{\mathbf{Z}}^{(t+1)}) + \mathbf{Q}^{-\frac{1}{2}} \boldsymbol{\xi}^{(t+1)}, \end{aligned} \quad (30)$$

where $\boldsymbol{\mu}(\cdot)$ and \mathbf{Q} are defined in (18). Note that $\mathbf{V}^{(t)}$ and $\tilde{\mathbf{V}}^{(t)}$ are distributed according to $\delta_{\mathbf{v}} P_{\rho, \gamma}^t$ and $\delta_{\tilde{\mathbf{v}}} P_{\rho, \gamma}^t$, respectively. Hence, by definition of the Wasserstein distance of order 2, it follows that

$$W_2(\delta_{\mathbf{v}} P_{\rho, \gamma}^t, \delta_{\tilde{\mathbf{v}}} P_{\rho, \gamma}^t) \leq \mathbb{E} \left[\|\mathbf{V}^{(t)} - \tilde{\mathbf{V}}^{(t)}\|^2 \right]^{1/2}. \quad (31)$$

In this section, our attention is directed towards establishing an upper limit for the squared norm $\|\mathbf{V}^{(t)} - \tilde{\mathbf{V}}^{(t)}\|^2$. This upper bound yields a specific limit on the Wasserstein distance, utilizing the preceding inequality. We begin by constraining (31) through the utilization of the subsequent technical lemma.

Lemma 1. *Let $\gamma \in \mathbb{R}_+^*$. If Assumptions 2-3 hold then, for any $\mathbf{v} = (\mathbf{x}, \mathbf{z})^\top$ and $\tilde{\mathbf{v}} = (\tilde{\mathbf{x}}, \tilde{\mathbf{z}})^\top$, for any $t \geq 1$,*

$$W_2(\delta_{\mathbf{v}} P_{\rho, \gamma}^t, \delta_{\tilde{\mathbf{v}}} P_{\rho, \gamma}^t) \leq \kappa_\gamma^{t-1} C_1^{1/2} \left[\kappa_\gamma \|\mathbf{z} - \tilde{\mathbf{z}}\| + \frac{\gamma}{\rho^2} \|\mathbf{x} - \tilde{\mathbf{x}}\| \right]$$

where $\kappa_\gamma = \max \{ |1 - \gamma \beta m_g|, |1 - \gamma(\beta M_g + 1/\rho^2)| \}$ and $C_1 = 1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2$.

Proof. Consider $(\mathbf{V}^{(t)}, \tilde{\mathbf{V}}^{(t)})_{t \in \mathbb{N}}$ defined in (27). Let $t \geq 0$. By (30), we have $\mathbf{X}^{(t+1)} - \tilde{\mathbf{X}}^{(t+1)} = \frac{1}{\rho^2} \mathbf{Q}^{-1} (\mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)})$ which implies

$$\begin{aligned} \|\mathbf{V}^{(t+1)} - \tilde{\mathbf{V}}^{(t+1)}\|^2 & \\ & \leq \left(1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2 \right) \|\mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)}\|^2. \end{aligned} \quad (32)$$

By (31) and 32, we need to bound $\|\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}\|^2$. By (29), we have

$$\begin{aligned} \mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)} &= \left(1 - \frac{\gamma}{\rho^2}\right) (\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}) + \frac{\gamma}{\rho^2} (\mathbf{X}^{(t)} - \tilde{\mathbf{X}}^{(t)}) \\ &\quad - \gamma \beta \left(\nabla g(\mathbf{Z}^{(t)}) - \nabla g(\tilde{\mathbf{Z}}^{(t)}) \right). \end{aligned}$$

Since $g(\cdot)$ is twice differentiable, we have

$$\begin{aligned} \nabla g(\mathbf{Z}^{(t)}) - \nabla g(\tilde{\mathbf{Z}}^{(t)}) &= \int_0^1 \nabla^2 g(\tilde{\mathbf{Z}}^{(t)} + t(\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)})) dt \\ &\quad \times (\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}). \end{aligned}$$

Using $\mathbf{X}^{(t)} - \tilde{\mathbf{X}}^{(t)} = \frac{1}{\rho^2} \mathbf{Q}^{-1} (\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)})$, it follows that

$$\begin{aligned} \mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)} &= \left(\left[1 - \frac{\gamma}{\rho^2}\right] \mathbf{I}_n + \frac{1}{\rho^4} \mathbf{Q}^{-1} \right) (\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}) \\ &\quad - \gamma \beta \int_0^1 \nabla^2 g(\tilde{\mathbf{Z}}^{(t)} + t(\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)})) dt \cdot (\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}). \end{aligned}$$

Let us define $\tilde{\mathbf{A}}^\top = \left[\frac{1}{\sigma} \mathbf{A}^\top \quad \frac{1}{\rho} \mathbf{I}_n \right]$. The precision matrix \mathbf{Q} can then be rewritten as $\mathbf{Q} = \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}}$. We define the orthogonal projector

$$\mathbf{P} = \tilde{\mathbf{A}} \mathbf{Q}^{-1} \tilde{\mathbf{A}}^\top. \quad (33)$$

By denoting $\mathbf{J} = [\mathbf{0}_n \quad \frac{1}{\rho} \mathbf{I}_n]^\top$, we have $\mathbf{J}^\top \mathbf{P} \mathbf{J} = \frac{1}{\rho^2} \mathbf{Q}^{-1}$. Considering

$$\mathbf{D}_g^{(t)} = \frac{\gamma}{\rho^2} \mathbf{I}_n + \gamma \beta \int_0^1 \nabla^2 g \left(\tilde{\mathbf{Z}}^{(t)} + s(\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}) \right) ds$$

and the projection matrix \mathbf{P} defined in (33), the difference $\mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)}$ can be rewritten as

$$\mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)} = \left(\mathbf{I}_n - \mathbf{D}_g^{(t)} + \gamma \mathbf{J}^\top \mathbf{P} \mathbf{J} \right) (\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}).$$

Under Assumptions 2-3 and exploiting the fact that \mathbf{P} is an orthogonal projector, thus $\mathbf{P} \preceq \mathbf{I}$, we have

$$\left[1 - \gamma(\beta M_g + \frac{1}{\rho^2}) \right] \mathbf{I}_n \preceq \mathbf{I}_n - \mathbf{D}_g^{(t)} + \gamma \mathbf{J}^\top \mathbf{P} \mathbf{J} \preceq \left[1 - \gamma \beta m_g \right] \mathbf{I}_n.$$

Therefore, we get $\|\mathbf{Z}^{(t+1)} - \tilde{\mathbf{Z}}^{(t+1)}\| \leq \kappa_\gamma \|\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}\|$. An immediate induction shows that, for all $t \geq 1$,

$$\|\mathbf{Z}^{(t)} - \tilde{\mathbf{Z}}^{(t)}\| \leq \kappa_\gamma^{t-1} \|\mathbf{Z}_1 - \tilde{\mathbf{Z}}_1\|. \quad (34)$$

In addition, we have $\mathbf{Z}_1 - \tilde{\mathbf{Z}}_1 = (\mathbf{I}_n - \mathbf{D}_g^{(0)})(\mathbf{z} - \tilde{\mathbf{z}}) + \frac{\gamma}{\rho^2}(\mathbf{x} - \tilde{\mathbf{x}})$. The triangle inequality gives

$$\begin{aligned} \|\mathbf{Z}_1 - \tilde{\mathbf{Z}}_1\| &\leq \|\mathbf{I}_n - \frac{\gamma}{\rho^2} \mathbf{I}_n - \mathbf{D}_g^{(0)}\| \|\mathbf{z} - \tilde{\mathbf{z}}\| + \frac{\gamma}{\rho^2} \|\mathbf{x} - \tilde{\mathbf{x}}\| \\ &\leq \kappa_\gamma \|\mathbf{z} - \tilde{\mathbf{z}}\| + \frac{\gamma}{\rho^2} \|\mathbf{x} - \tilde{\mathbf{x}}\|. \end{aligned}$$

Combining (34) and the previous inequality and using 32, we get for $t \geq 1$,

$$\begin{aligned} \|\mathbf{V}^{(t)} - \tilde{\mathbf{V}}^{(t)}\|^2 &\leq \kappa_\gamma^{2(t-1)} \left(1 + \left\| \frac{1}{\rho^2} \mathbf{Q}^{-1} \right\|^2 \right) \\ &\quad \times \left(\kappa_\gamma \|\mathbf{z} - \tilde{\mathbf{z}}\| + \frac{\gamma}{\rho^2} \|\mathbf{x} - \tilde{\mathbf{x}}\| \right)^2. \end{aligned}$$

The proof is concluded by (31). \square

On the basis of the preceding lemma, we provide in what follows the proof of Proposition 2.

Proof. Note that the condition $0 < \gamma \leq 2(\beta m_g + \beta M_g + 1/\rho^2)^{-1}$ ensures that $\kappa_\gamma = 1 - \gamma \beta m_g \in (0, 1)$. From Lemma 1 combined with [41, Lemma 20.3.2, Theorem 20.3.4], $P_{\rho, \delta}$ admits a unique invariant probability distribution $\pi_{\rho, \gamma}$. Moreover, for any $\mathbf{v} = (\mathbf{x}, \mathbf{z})^\top$ with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{z} \in \mathbb{R}^n$ and any $t \in \mathbb{N}^*$, we have

$$\begin{aligned} W_2^2(\delta_{\mathbf{v}} P_{\rho, \gamma}^t, \pi_{\rho, \gamma}) &\leq \kappa_\gamma^{2(t-1)} \left(1 + \left\| \frac{1}{\rho^2} \mathbf{Q}^{-1} \right\|^2 \right) \\ &\quad \times \int_{\mathbb{R}^n \times \mathbb{R}^n} \left[(1 - \gamma \beta m_g) \|\mathbf{z} - \tilde{\mathbf{z}}\| + \frac{\gamma}{\rho^2} \|\mathbf{x} - \tilde{\mathbf{x}}\| \right]^2 d\pi_{\rho, \gamma}(\tilde{\mathbf{v}}). \end{aligned}$$

\square

APPENDIX C PROOF OF PROPOSITION 3

In this section, we establish an explicit bound on $W_2^2(\pi_{\rho, \gamma}, \pi_\rho)$ where π_ρ is the target augmented distribution. Consider first for any $\mathbf{x} \in \mathbb{R}^n$, the stochastic differential equation (SDE) defined by

$$d\tilde{\mathbf{Y}}_t^{\mathbf{x}} = -\nabla U(\tilde{\mathbf{Y}}_t^{\mathbf{x}}) dt - \frac{1}{\rho^2} \mathbf{x} + \sqrt{2} d\mathbf{B}_t \quad (35)$$

where $(\mathbf{B}_t)_{t \geq 0}$ is a n -dimensional Brownian motion and the potential $U_{\mathbf{x}}(\cdot)$ is defined as

$$U_{\mathbf{x}}(\cdot) = \beta g(\cdot) + \frac{1}{2\rho^2} \|\cdot - \mathbf{x}\|^2 \quad (36)$$

and, to lighten the notations, we denote the potential in 35, $U(\cdot) = U_0(\cdot)$. Note that under Assumption 2, this SDE admits a unique strong solution [42, Chapter IX, Theorem (2.1)]. Denote the Markov semi-group associated to (35) by $(\tilde{R}_{\rho, t})_{t \geq 0}$ defined for any $\tilde{\mathbf{y}}_0 \in \mathbb{R}^n$, $t \geq 0$ and $\mathbf{B} \in \mathcal{B}(\mathbb{R}^n)$ by $\tilde{R}_{\rho, t}(\tilde{\mathbf{y}}_0, \mathbf{B} | \mathbf{x}) = \mathbb{P}(\tilde{\mathbf{Y}}_t^{\mathbf{x}, \tilde{\mathbf{y}}_0} \in \mathbf{B})$, where $(\tilde{\mathbf{Y}}_t^{\mathbf{x}, \tilde{\mathbf{y}}_0})_{t \geq 0}$ is a solution of (35) with $\tilde{\mathbf{Y}}_0^{\mathbf{x}, \tilde{\mathbf{y}}_0} = \tilde{\mathbf{y}}_0$. We consider the Markov kernel defined, for any $\mathbf{v} = (\mathbf{x}, \mathbf{z})^\top$ and $\mathbf{A} \in \mathcal{B}(\mathbb{R}^n)$, $\mathbf{B} \in \mathcal{B}(\mathbb{R}^n)$, by

$$\tilde{P}_{\rho, \gamma}(\mathbf{v}, \mathbf{A} \times \mathbf{B}) = \int_{\mathbf{B}} \tilde{R}_{\rho, \gamma}(\mathbf{z}, d\tilde{\mathbf{z}} | \mathbf{x}) \int_{\mathbf{A}} p(d\mathbf{x} | \mathbf{y}, \tilde{\mathbf{z}}). \quad (37)$$

Note that $P_{\rho, \gamma}$ can be interpreted as a discretized version of $\tilde{P}_{\rho, \gamma}$ using the Euler-Maruyama scheme. Under Assumption 2, the Markov kernel $\tilde{P}_{\rho, \gamma}$ defined by (37) admits π_ρ as an invariant probability distribution [34, Proposition S21], i.e., $\forall t \geq 0$, $\pi_\rho \tilde{P}_{\rho, \gamma}^t = \pi_\rho$.

Let $(\mathbf{B}_t)_{t \geq 0}$ an i.i.d. n -dimensional Brownian motion and let $(\boldsymbol{\xi}^{(t)})_{t \geq 0}$ be a sequence of i.i.d. standard n -dimensional Gaussian random variables independent of $(\mathbf{B}_t)_{t \geq 0}$. For $t \geq 0$, we define the synchronous coupling $\mathbf{V}^{(t)} = (\mathbf{X}^{(t)}, \mathbf{Z}^{(t)})$ and $(\tilde{\mathbf{V}}^{(t)} = (\tilde{\mathbf{X}}^{(t)}, \tilde{\mathbf{Z}}^{(t)})$, starting from $(\mathbf{X}^{(0)}, \mathbf{Z}^{(0)}) = (\mathbf{x}, \mathbf{z})$, $(\tilde{\mathbf{X}}^{(0)}, \tilde{\mathbf{Z}}^{(0)})$ distributed according to π_ρ

$$\begin{aligned} \tilde{\mathbf{Z}}^{(t+1)} &= \tilde{\mathbf{Y}}_{(t+1)\gamma}, & \tilde{\mathbf{X}}^{(t+1)} &= \boldsymbol{\mu}(\tilde{\mathbf{Z}}^{(t+1)}) + \mathbf{Q}^{-\frac{1}{2}} \boldsymbol{\xi}^{(t+1)}, \\ \mathbf{Z}^{(t+1)} &= \mathbf{Y}_{(t+1)\gamma}, & \mathbf{X}^{(t+1)} &= \boldsymbol{\mu}(\mathbf{Z}^{(t+1)}) + \mathbf{Q}^{-\frac{1}{2}} \boldsymbol{\xi}^{(t+1)}, \end{aligned} \quad (38)$$

with

$$\begin{aligned} \tilde{\mathbf{Y}}_{(t+1)\gamma} &= \tilde{\mathbf{Y}}_{t\gamma} - \int_{t\gamma}^{(t+1)\gamma} \nabla U(\tilde{\mathbf{Y}}_l) dl \\ &\quad + \frac{\gamma}{\rho^2} \tilde{\mathbf{X}}^{(t)} + \sqrt{2}(\mathbf{B}_{(t+1)\gamma} - \mathbf{B}_{t\gamma}) \\ \mathbf{Y}_{(t+1)\gamma} &= \mathbf{Y}_{t\gamma} - \gamma \nabla U(\mathbf{Y}_{t\gamma}) \\ &\quad + \frac{\gamma}{\rho^2} \mathbf{X}^{(t)} + \sqrt{2}(\mathbf{B}_{(t+1)\gamma} - \mathbf{B}_{t\gamma}). \end{aligned} \quad (39)$$

The stochastic processes $(\mathbf{V}^{(t)}, \tilde{\mathbf{V}}^{(t)})_{t \geq 0}$ satisfy (32). Note that $\mathbf{V}^{(t)}$ and $\tilde{\mathbf{V}}^{(t)}$ are distributed according to $\pi_\rho \tilde{P}_{\rho, \gamma}^t = \pi_\rho$ and $\delta_{\tilde{\mathbf{v}}} P_{\rho, \gamma}^t$, respectively. Hence, by definition of the Wasserstein distance of order 2, it follows that

$$W_2(\pi_\rho, \delta_{\tilde{\mathbf{v}}} P_{\rho, \gamma}^t) \leq \mathbb{E} \left[\|\mathbf{V}^{(t)} - \tilde{\mathbf{V}}^{(t)}\|^2 \right]^{1/2}. \quad (40)$$

We start this section by a first estimate on $\mathbb{E} \left[\|\mathbf{V}^{(t)} - \tilde{\mathbf{V}}^{(t)}\|^2 \right]^{1/2}$. The following result holds regarding the process $(\tilde{\mathbf{Y}}_t)_{t \in \mathbb{N}_+}$ defined in (39).

Lemma 2. *If Assumptions 2 and 3 hold, let define $\tilde{M} = \beta M_g + 1/\rho^2$ and $\gamma \in \mathbb{R}_+^*$ such that $\gamma < 1/\tilde{M}$. Then, for any $t \geq 1$,*

$$\tilde{\mathbf{Z}}^{(t+1)} - \mathbf{Z}^{(t+1)} = \mathbf{T}_1^{(t)}(\tilde{\mathbf{Z}}^{(t)} - \mathbf{Z}^{(t)}) - \mathbf{T}_2^{(t)}, \quad (41)$$

where $(\mathbf{Z}^{(t)}, \tilde{\mathbf{Z}}^{(t)})_{t \in \mathbb{N}}$ is defined in (38), $\mathbf{T}_1^{(t)}$ and $\mathbf{T}_2^{(t)}$ are given by

$$\mathbf{T}_1^{(t)} = \mathbf{I}_n + \gamma \mathbf{J}^\top \mathbf{P} \mathbf{J} - \gamma \int_0^1 \nabla^2 U((1-s)\mathbf{Y}_{t\gamma} + s\tilde{\mathbf{Y}}_{t\gamma}) ds \quad (42)$$

$$\mathbf{T}_2^{(t)} = \int_0^\gamma [\nabla U(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U(\tilde{\mathbf{Y}}_{t\gamma})] dl. \quad (43)$$

Proof. Since $U(\cdot)$ is twice differentiable, we have

$$\begin{aligned} & \nabla U(\tilde{\mathbf{Y}}_{t\gamma}) - \nabla U(\mathbf{Y}_{t\gamma}) \\ &= \left(\int_0^1 \nabla^2 U((1-s)\mathbf{Y}_{t\gamma} + s\tilde{\mathbf{Y}}_{t\gamma}) ds \right) \times (\tilde{\mathbf{Y}}_{t\gamma} - \mathbf{Y}_{t\gamma}). \end{aligned}$$

By using $\frac{1}{\rho^2}(\tilde{\mathbf{X}}^{(t)} - \mathbf{X}^{(t)}) = \mathbf{J}^\top \mathbf{P} \mathbf{J}(\tilde{\mathbf{Y}}_{t\gamma} - \mathbf{Y}_{t\gamma})$, it follows from (39) that

$$\begin{aligned} & \tilde{\mathbf{Y}}_{(t+1)\gamma} - \mathbf{Y}_{(t+1)\gamma} = (\tilde{\mathbf{Y}}_{t\gamma} - \mathbf{Y}_{t\gamma}) \\ & \times \left(\mathbf{I}_n - \gamma \int_0^1 \nabla^2 U((1-s)\mathbf{Y}_{t\gamma} + s\tilde{\mathbf{Y}}_{t\gamma}) ds + \gamma \mathbf{J}^\top \mathbf{P} \mathbf{J} \right) \\ & - \int_0^\gamma [\nabla U(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U(\tilde{\mathbf{Y}}_{t\gamma})] dl. \end{aligned}$$

where \mathbf{P} is defined in (33). By (38), we have

$$\tilde{\mathbf{Z}}^{(t+1)} - \mathbf{Z}^{(t+1)} = \mathbf{T}_1^{(t)}(\tilde{\mathbf{Z}}^{(t)} - \mathbf{Z}^{(t)}) - \mathbf{T}_2^{(t)}.$$

□

Based on Lemma 2, we have the following relation between $\|\tilde{\mathbf{Z}}^{(t+1)} - \mathbf{Z}^{(t+1)}\|^2$ and $\|\tilde{\mathbf{Z}}^{(t)} - \mathbf{Z}^{(t)}\|^2$.

Lemma 3. *If Assumption 3 holds, let define $\tilde{M} = \beta M_g + 1/\rho^2$ and $\gamma \in \mathbb{R}_+^*$ such that $\gamma < 1/\tilde{M}$. Then, for any $\epsilon > 0$ and $t \geq 1$,*

$$\begin{aligned} & \|\tilde{\mathbf{Z}}^{(t+1)} - \mathbf{Z}^{(t+1)}\|^2 \\ & \leq (1 + 2\epsilon) \|\mathbf{T}_1^{(t)}\|^2 \|\tilde{\mathbf{Z}}^{(t)} - \mathbf{Z}^{(t)}\|^2 + (1 + \frac{1}{2\epsilon}) \|\mathbf{T}_2^{(t)}\|^2. \end{aligned}$$

where $(\mathbf{Z}^{(t)}, \tilde{\mathbf{Z}}^{(t)})_{t \in \mathbb{N}}$ is defined in (38).

Proof. The proof of Lemma (3) can be found in [34]. □

We now have the following result regarding the contracting term.

Lemma 4. *Let define $\tilde{M} = \beta M_g + 1/\rho^2$ and $\gamma \in \mathbb{R}_+^*$ such that $\gamma < 1/\tilde{M}$. If assumptions 2 and 3 hold then, for any $n \geq 0$,*

$$\|\mathbf{T}_1^{(t)}\| \leq 1 - \gamma\beta m_g,$$

where $\mathbf{T}_1^{(t)}$ is defined in (42).

Proof. By denoting

$$\begin{aligned} \mathbf{C}^{(t)} &= \gamma \int_0^1 \nabla^2 U((1-s)\mathbf{Y}_{t\gamma} + s\tilde{\mathbf{Y}}_{t\gamma}) ds \\ &= \frac{\gamma}{\rho^2} \mathbf{I}_n + \gamma \int_0^1 \nabla^2 g((1-s)\mathbf{Y}_{t\gamma} + s\tilde{\mathbf{Y}}_{t\gamma}) ds \end{aligned}$$

Assumptions 2 and 3 imply $(\frac{\gamma}{\rho^2} + \gamma\beta m_g)\mathbf{I}_n \preceq \mathbf{C}^{(t)} \preceq (\frac{\gamma}{\rho^2} + 2\gamma\beta)\mathbf{I}_n$. Since \mathbf{P} is an orthogonal projection (33), $\mathbf{P} \preceq \mathbf{I}$ and

$\mathbf{0}_n \preceq \gamma \mathbf{J}^\top \mathbf{P} \mathbf{J} \preceq \frac{\gamma}{\rho^2} \mathbf{I}_n$. Subtracting these previous inequalities and adding \mathbf{I}_n leads to

$$(1 - \gamma\tilde{M})\mathbf{I}_n \preceq \mathbf{I}_n - \mathbf{C}^{(t)} + \gamma \mathbf{J}^\top \mathbf{P} \mathbf{J} \preceq (1 - \gamma\beta m_g)\mathbf{I}_n$$

and

$$\|\mathbf{T}_1^{(t)}\| \leq \max(|1 - \gamma\beta m_g|, |1 - \gamma\tilde{M}|) = 1 - \gamma\beta m_g. \quad \square$$

The following lemma provides an upper bound on $\|\mathbf{T}_2^{(t)}\|^2$.

Lemma 5. *If Assumptions 2 and 3 hold, let define $\tilde{m} = \beta m_g + 1/\rho^2$, $\tilde{M} = \beta M_g + 1/\rho^2$ and $\gamma \in \mathbb{R}_+^*$ such that $\gamma < 1/\tilde{M}$. Then, for any $t \in \mathbb{N}$,*

$$\mathbb{E} \left[\|\mathbf{T}_2^{(t)}\|^2 \right] \leq n\gamma^2 \tilde{M}^2 \left[1 + \frac{\gamma^2 \tilde{M}^2}{12} + \frac{\gamma \tilde{M}^2}{2\tilde{m}} \right],$$

where $\mathbf{T}_2^{(t)}$ is defined in (43).

Proof. Let $t \in \mathbb{N}$. We have

$$\|\mathbf{T}_2^{(t)}\|^2 = \left\| \int_0^\gamma [\nabla U(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U(\tilde{\mathbf{Y}}_{t\gamma})] dl \right\|^2. \quad (44)$$

With the previous result and the Jensen inequality,

$$\|\mathbf{T}_2^{(t)}\|^2 \leq \int_0^\gamma \left\| \nabla U(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U(\tilde{\mathbf{Y}}_{t\gamma}) \right\|^2 dl. \quad (45)$$

Let denote $\mathcal{G}_0 = \sigma(\mathbf{Z}^{(0)}, \tilde{\mathbf{Z}}^{(0)}, \mathbf{X}^{(0)}, \tilde{\mathbf{X}}^{(0)})$, for any $t \in \mathbb{N}^*$, $\mathcal{G}_t = \sigma\{\mathbf{Z}^{(0)}, \tilde{\mathbf{Z}}^{(0)}, \mathbf{X}^{(0)}, \tilde{\mathbf{X}}^{(0)}, (\mathbf{B}_k)_{k \geq 0}, k \leq t\}$ and \mathcal{F}_t the σ -field generated by \mathcal{G}_{t-1} . Using [26, Lemma 21] applied to the potential (36) yields

$$\begin{aligned} & \int_0^\gamma \mathbb{E}^{\mathcal{F}_{t\gamma}} \|\nabla U(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U(\tilde{\mathbf{Y}}_{t\gamma})\|^2 dl \\ &= \int_0^\gamma \mathbb{E}^{\mathcal{F}_{t\gamma}} \|\nabla U_{\tilde{\mathbf{x}}^{(t)}}(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U_{\tilde{\mathbf{x}}^{(t)}}(\tilde{\mathbf{Y}}_{t\gamma})\|^2 dl \\ &\leq \gamma^2 \tilde{M}^2 \left[n + \frac{n\gamma^2 \tilde{M}^2}{12} + \frac{\gamma \tilde{M}^2}{2} \|\tilde{\mathbf{Y}}_{t\gamma} - \mathbf{z}_{t,*}\|^2 \right], \quad (46) \end{aligned}$$

where $\mathbf{z}_{t,*} = \arg \min_{\mathbf{z} \in \mathbb{R}^n} U_{\tilde{\mathbf{x}}^{(t)}}(\mathbf{z})$.

By (46) and using [26, Proposition 1], we get

$$\begin{aligned} & \int_0^\gamma \mathbb{E} \|\nabla U(\tilde{\mathbf{Y}}_{t\gamma+l}) - \nabla U(\tilde{\mathbf{Y}}_{t\gamma})\|^2 dl \leq n\gamma^2 \tilde{M}^2 \\ & \quad \times \left[1 + \frac{\gamma^2 \tilde{M}^2}{12} + \frac{\gamma \tilde{M}^2}{2\tilde{m}} \right]. \end{aligned}$$

Combining this result with (45) completes the proof. □

We can now combine Lemma 5 and Lemma 4 with Lemma 2 to get the following bound.

Lemma 6. *If Assumptions 2 and 3 hold, let denote $\tilde{m} = \beta m_g + 1/\rho^2$, $\tilde{M} = \beta M_g + 1/\rho^2$, $\gamma \in \mathbb{R}_+^*$ such that $\gamma < 1/\tilde{M}$ and $r_\gamma = \gamma\beta m_g \in (0, 1)$. Then, for $t \geq 1$, we have*

$$\begin{aligned} \mathbb{E} \left[\|\tilde{\mathbf{Z}}^{(t)} - \mathbf{Z}^{(t)}\|^2 \right] &\leq (1 - r_\gamma + r_\gamma^2/2)^{2(t-1)} \mathbb{E} \left[\|\tilde{\mathbf{Z}}^{(1)} - \mathbf{Z}^{(1)}\|^2 \right] \\ &\quad + 2(\beta m_g)^{-1} \times n\gamma \tilde{M}^2 \left(1 + \frac{\gamma \tilde{M}^2}{12} + \frac{\gamma \tilde{M}^2}{2\tilde{m}} \right). \end{aligned}$$

Proof. The proof of Lemma (6) can be found in [34]. □

Lemma 7. *If Assumption 3 holds, let denote $\tilde{m} = \beta m_g + 1/\rho^2$, $\tilde{M} = \beta M_g + 1/\rho^2$, $\gamma \in \mathbb{R}_+^*$ such that $\gamma < 1/\tilde{M}$ and $r_\gamma \in (0, 1)$. Then, for any $\mathbf{v} \in \mathbb{R}^{n \times n}$ and $t \geq 1$, the following holds*

$$\begin{aligned} W_2^2(\delta_{\mathbf{v}} P_{\rho, \gamma}^t, \pi_\rho) &\leq (1 - r_\gamma + r_\gamma^2/2)^{2(t-1)} (1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2) \\ &\quad \times \mathbb{E} \left[\|\tilde{\mathbf{Z}}^{(1)} - \mathbf{Z}^{(1)}\|^2 \right] + \frac{2(1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2)}{\beta m_g} \\ &\quad \times n\gamma \tilde{M}^2 \left[1 + \frac{\gamma^2 \tilde{M}^2}{12} + \gamma \tilde{M}^2 / (2\tilde{m}) \right], \end{aligned}$$

where $P_{\rho, \gamma}$ is defined in (26) and $(\tilde{\mathbf{Z}}^{(t)}, \mathbf{Z}^{(t)})_{t \in \mathbb{N}}$ is defined in (38).

Proof. The proof follows from the combination of (40) with (32) and Lemma 6. \square

Now we can give the proof of Proposition 3.

Proof. Note that the condition $\gamma \leq 2(\beta m_g + \beta M_g + 1/\rho^2)^{-1}$ ensures that $1 - r_\gamma + r_\gamma^2/2 \in (0, 1)$. By Proposition 2, $\delta_{\mathbf{v}} P_{\rho, \gamma}^t$ converges in W_2 to $\pi_{\rho, \gamma}$. Therefore, using Lemma 7 and taking $t \rightarrow \infty$, we obtain

$$\begin{aligned} W_2^2(\pi_\rho, \pi_{\rho, \gamma}) &\leq \frac{2(1 + \frac{1}{\rho^2} \|\mathbf{Q}^{-1}\|^2)}{\beta m_g} \\ &\quad \times n\gamma \tilde{M}^2 \left(1 + \frac{\gamma^2 \tilde{M}^2}{12} + \frac{\gamma \tilde{M}^2}{2\tilde{m}} \right). \end{aligned}$$

\square

APPENDIX D

AUGMENTED DISTRIBUTION AND RED-LWSGS FOR SUPER-RESOLUTION

This appendix details the AXDA model and the corresponding sampling algorithm when tackling the super-resolution task. In this case, the operator \mathbf{A} can be written as $\mathbf{A} = \mathbf{S}\mathbf{B}$ where $\mathbf{B} \in \mathbb{R}^{n \times n}$ is a circulant matrix standing for a spatially invariant blur and $\mathbf{S} \in \mathbb{R}^{m \times n}$ stands for a regular downsampling operator. When directly adopting the splitting trick proposed in Section III-A, sampling according to the conditional distribution (15) remains difficult because the precision matrix is neither diagonal (as for the inpainting task) nor diagonalizable in the Fourier domain (as for the deblurring task). To overcome this difficulty, one suitable AXDA consists in introducing two splitting variables, which allows the operators \mathbf{B} and \mathbf{S} to be decoupled. This leads to the augmented posterior distribution

$$\begin{aligned} \pi_{\rho_1, \rho_2}(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2) &\propto \exp \left[-\frac{1}{2\sigma^2} \|\mathbf{S}\mathbf{z}_1 - \mathbf{y}\|_2^2 - \beta g_{\text{red}}(\mathbf{z}_2) \right. \\ &\quad \left. - \frac{1}{2\rho_1^2} \|\mathbf{B}\mathbf{x} - \mathbf{z}_1\|_2^2 - \frac{1}{2\rho_2^2} \|\mathbf{x} - \mathbf{z}_2\|_2^2 \right]. \end{aligned}$$

The associated SGS alternatively samples according to the three conditional distributions

$$p(\mathbf{z}_1 | \mathbf{x}, \mathbf{y}) \propto \exp \left[-\frac{1}{2\sigma^2} \|\mathbf{S}\mathbf{z}_1 - \mathbf{y}\|_2^2 - \frac{1}{2\rho_1^2} \|\mathbf{B}\mathbf{x} - \mathbf{z}_1\|_2^2 \right] \quad (47)$$

$$p(\mathbf{x} | \mathbf{z}_1, \mathbf{z}_2) \propto \exp \left[-\frac{1}{2\rho_1^2} \|\mathbf{B}\mathbf{x} - \mathbf{z}_1\|_2^2 - \frac{1}{2\rho_2^2} \|\mathbf{x} - \mathbf{z}_2\|_2^2 \right] \quad (48)$$

$$p(\mathbf{z}_2 | \mathbf{x}) \propto \exp \left[-\frac{1}{2} \mathbf{x}^\top (\mathbf{z}_2 - \mathbf{D}_\nu(\mathbf{z}_2)) - \frac{1}{2\rho_2^2} \|\mathbf{x} - \mathbf{z}_2\|_2^2 \right]$$

It appears that (47) and (48) define the conditional posteriors associated with the inpainting and deblurring tasks, respectively.

APPENDIX E

EXPERIMENTAL DETAILS

A. Pretrained denoiser

All experiments have been performed with DRUNet as the pre-trained denoiser used by the PnP- and RED-based methods. This denoiser $\mathbf{D}_\nu(\cdot)$ has the ability to handle different noise levels with a single model thanks to the parameter ν which controls the strength of the denoising. This parameter has been estimated following the strategy recommended in [11]. The parameter ν is uniformly sampled from a large noise level $\nu^{(1)}$ to a small one $\nu^{(N_{\text{bi}})}$ according to a logarithmic scale, which results in a sequence of $\nu^{(1)} > \nu^{(2)} > \dots > \nu^{(N_{\text{bi}})}$. Following [11], $\nu^{(1)}$ is fixed to 49 while $\nu^{(N_{\text{bi}})}$ is adjusted w.r.t. the image noise level σ . For the sampling-based methods, to ensure the stationary of the kernel, the noise parameter is set to a fixed value beyond the burn-in period, i.e., $\forall t \in [N_{\text{bi}}, N_{\text{MC}}]$, $\nu^{(t)} = \nu^{(N_{\text{bi}})}$.

B. Implementation details regarding RED-LwSGS

This appendix provides additional details regarding the implementation of the proposed LwSGS algorithm. The regularization parameter and the coupling parameters have been adjusted to reach the best performance. For experiments on the FFHQ data set, the regularization parameter β is set to 8.0×10^{-2} , 1.25×10^{-1} and 1.0 for the deblurring, inpainting and super-resolution tasks, respectively, while it has been fixed to 4.89×10^{-3} , 1.167×10^{-1} and 4.966×10^{-2} for the experiments conducted on the ImageNet data set. The other parameters are fixed as $(N_{\text{MC}}, N_{\text{bi}}, \rho^2, \gamma) = (5000, 2000, 6 \times 10^{-8}, \frac{0.99}{2\beta+1/\rho^2})$ for deblurring, $(N_{\text{MC}}, N_{\text{bi}}, \rho^2, \gamma) = (10000, 4500, 1.5, \frac{0.99}{2\beta+1/\rho^2})$ for inpainting and $(N_{\text{MC}}, N_{\text{bi}}, \rho_1^2, \rho_2^2, \gamma) = (12500, 3500, 2 \times 10^{-1}, 1, \frac{0.8}{2\beta+1/\rho_2^2})$ for super-resolution which requires a double splitting.

C. Implementation details regarding the compared methods

This appendix provides additional details regarding the implementation of the compared methods. First, RED-ADMM, RED-HQS, PnP-ADMM, PnP-ULA and Diff-PIR are implemented using the same denoiser as the proposed method (see Appendix E-A). For the sampling-based methods, i.e., PnP-ULA, TV-MYULA, TV-SP, the total number of iterations have been set as for the proposed RED-LwSGS (see Appendix E-B).

For optimization-based algorithms, i.e. RED-ADMM, RED-HQS and PnP-ADMM, the total number of iterations is set as follows: 150 for deblurring and 350 for inpainting and super-resolution. Finally all model and algorithmic parameters have been adjusted to reach the best performance. More precisely, for the experiments conducted on the FFHQ data:

- RED-ADMM: the hyperparameters (α, λ, β) have been set to $(2, 2 \times 10^{-3}, 9 \times 10^{-4})$ for deblurring, $(2, 10^{-2}, 4 \times 10^{-2})$ for inpainting and $(2, 8 \times 10^{-3}, 10^{-6})$ for super-resolution.
- RED-HQS: similarly to RED-ADMM, the hyperparameters are $(\alpha, \lambda, \beta) = (2, 10^{-2}, 4 \times 10^{-3})$ for deblurring, $(\alpha, \lambda, \beta) = (2, 2 \times 10^{-2}, 1.8 \times 10^{-2})$ for inpainting and $(\alpha, \lambda, \beta) = (2, 8 \times 10^{-3}, 10^{-6})$ for super-resolution.
- PnP-ADMM: the parameter ρ is set to 10^{-4} for deblurring, 10^{-3} for inpainting and 6.5×10^{-2} for super-resolution.
- PnP-ULA: the parameters (N_{bi}, β) have been set to $(2500, 7.3 \times 10^{-4})$ for deblurring, $(5000, 10^{-4})$ for inpainting and $(6000, 2.75 \times 10^{-4})$ for super-resolution.
- TV-MYULA: the parameters (N_{bi}, β) have been set to $(2500, 10^{-1})$ for deblurring, $(5000, 8 \times 10^{-8})$ for inpainting and $(8500, 1)$ for super-resolution.
- DiffPIR: the parameters (λ, ζ) have been set to $(2, 3 \times 10^{-1})$ for deblurring and $(1, 1)$ for inpainting and super-resolution.
- TV-SP: the parameters $(N_{\text{bi}}, \rho, \beta)$ have been set to $(2000, 9 \times 10^{-4}, 1)$ for deblurring and $(4500, 1, 3.5 \times 10^{-3})$ for inpainting while for super-resolution, which requires a double splitting, the parameters $(N_{\text{bi}}, \rho_1^2, \rho_2^2, \beta)$ have been set to $(3500, 10^{-5}, 10^{-5}, 3.76)$.

REFERENCES

- [1] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.
- [2] W. C. Karl, “Regularization in image restoration and reconstruction,” in *Handbook of image and video processing*. Elsevier, 2005, pp. 183–V.
- [3] F. Cao, M. Cai, Y. Tan, and J. Zhao, “Image super-resolution via adaptive ℓ_p ($0 < p < 1$) regularization and sparse representation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1550–1561, 2016.
- [4] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, “An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems,” *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, 2010.
- [5] S. V. Venkatakrisnan, C. A. Bouman, and B. Wohlberg, “Plug-and-play priors for model based reconstruction,” in *Proc. IEEE Global Conf. Signal Info. Process. (GlobalSIP)*. IEEE, 2013, pp. 945–948.
- [6] D. Geman and C. Yang, “Nonlinear image recovery with half-quadratic regularization,” *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, 1995.
- [7] J. Douglas and H. H. Rachford, “On the numerical solution of heat conduction problems in two and three space variables,” *Trans. Am. Math. Soc.*, vol. 82, no. 2, pp. 421–439, 1956.
- [8] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *Proc. Int. Conf. Computer Vision Pattern Recognition (CVPR)*, vol. 2. Ieee, 2005, pp. 60–65.
- [9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.
- [10] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [11] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, “Plug-and-play image restoration with deep denoiser prior,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6360–6376, 2021.
- [12] S. H. Chan, X. Wang, and O. A. Elgendy, “Plug-and-play ADMM for image restoration: Fixed-point convergence and applications,” *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 84–98, 2016.
- [13] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, “Plug-and-play methods provably converge with properly trained denoisers,” in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2019, pp. 5546–5557.
- [14] S. Hurault, A. Leclaire, and N. Papadakis, “Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization,” in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2022, pp. 9483–9505.
- [15] Y. Romano, M. Elad, and P. Milanfar, “The little engine that could: Regularization by denoising (RED),” *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.
- [16] E. T. Reehorst and P. Schniter, “Regularization by denoising: Clarifications and new interpretations,” *IEEE Trans. Comput. Imag.*, vol. 5, no. 1, pp. 52–67, 2018.
- [17] R. Cohen, M. Elad, and P. Milanfar, “Regularization by denoising via fixed-point projection (RED-PRO),” *SIAM J. Imag. Sci.*, vol. 14, no. 3, pp. 1374–1406, 2021.
- [18] J. M. Bardsley, *Computational Uncertainty Quantification for Inverse Problems: An Introduction to Singular Integrals*. SIAM, 2018.
- [19] X. Cai, M. Pereyra, and J. D. McEwen, “Uncertainty quantification for radio interferometric imaging – I. Proximal MCMC methods,” *Monthly Notices of the Royal Astronomical Society*, vol. 480, no. 3, pp. 4154–4169, 2018.
- [20] M. Holden, M. Pereyra, and K. C. Zygalakis, “Bayesian imaging with data-driven priors encoded by neural networks,” *SIAM J. Imag. Sci.*, vol. 15, no. 2, pp. 892–924, 2022.
- [21] Z. Cai, J. Tang, S. Mukherjee, J. Li, C. B. Schönlieb, and X. Zhang, “NF-ULA: Langevin Monte Carlo with normalizing flow prior for imaging inverse problems,” *SIAM J. Imag. Sci.*, 2024.
- [22] R. Laumont, V. D. Bortoli, A. Almansa, J. Delon, A. Durmus, and M. Pereyra, “Bayesian imaging using plug & play priors: when Langevin meets Tweedie,” *SIAM J. Imag. Sci.*, vol. 15, no. 2, pp. 701–737, 2022.
- [23] M. Vono, N. Dobigeon, and P. Chainais, “Asymptotically exact data augmentation: Models, properties, and algorithms,” *J. Comput. Graph. Stat.*, vol. 30, no. 2, pp. 335–348, 2020.
- [24] M. Vono, N. Dobigeon, and P. Chainais, “Split-and-augmented Gibbs sampler – Application to large-scale inference problems,” *IEEE Trans. Signal Process.*, vol. 67, no. 6, pp. 1648–1661, 2019.
- [25] B. Efron, “Tweedie’s formula and selection bias,” *J. Amer. Stat. Soc.*, vol. 106, no. 496, pp. 1602–1614, 2011.
- [26] A. Durmus and E. Moulines, “High-dimensional Bayesian inference via the unadjusted Langevin algorithm,” *Bernoulli*, vol. 25, no. 4A, pp. 2854–2882, 2019.
- [27] L. J. Rendell, A. M. Johansen, A. Lee, and N. Whiteley, “Global consensus monte carlo,” *J. Comput. Graph. Stat.*, vol. 30, no. 2, pp. 249–259, 2020.
- [28] M. Vono, N. Dobigeon, and P. Chainais, “High-dimensional Gaussian sampling: a review and a unifying approach based on a stochastic proximal point algorithm,” *SIAM Rev.*, vol. 64, no. 1, pp. 3–56, 2022.
- [29] M. Vono, N. Dobigeon, and P. Chainais, “Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler,” in *Proc. IEEE Workshop Mach. Learning for Signal Process. (MLSP)*, Aalborg, Denmark, Sept. 2018.
- [30] M. Vono, N. Dobigeon, and P. Chainais, “Bayesian image restoration under Poisson noise and log-concave prior,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Brighton, U.K., April 2019.
- [31] G. O. Roberts and O. Stramer, “Langevin diffusions and Metropolis-Hastings algorithms,” *Methodology and computing in applied probability*, vol. 4, pp. 337–357, 2002.
- [32] P.-A. Thouvenin, A. Repetti, and P. Chainais, “A distributed Gibbs sampler with hypergraph structure for high-dimensional inverse problems,” *J. Comput. Graph. Stat.*, 2024.
- [33] A. Durmus, S. Majewski, and B. Miasojedow, “Analysis of Langevin Monte Carlo via convex optimization,” *J. Mach. Learning Research*, vol. 20, no. 1, pp. 2666–2711, 2019.
- [34] V. Plassier, M. Vono, A. Durmus, and E. Moulines, “DG-LMC: a turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs,” in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2021, pp. 8577–8587.
- [35] M. Terris, A. Repetti, J.-C. Pesquet, and Y. Wiaux, “Building firmly nonexpansive convolutional neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2020, pp. 8658–8662.

- [36] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. Int. Conf. Computer Vision Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
- [37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Int. Conf. Computer Vision Pattern Recognition (CVPR)*, 2009, pp. 248–255.
- [38] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau," *SIAM J. Imag. Sci.*, vol. 11, no. 1, pp. 473–506, 2018.
- [39] Y. Zhu, K. Zhang, J. Liang, J. Cao, B. Wen, R. Timofte, and L. V. Gool, "Denoising diffusion models for plug-and-play image restoration," in *Int. Conf. Computer Vision Pattern Recognition Workshops (NTIRE)*, 2023.
- [40] M. D. Fall and É. Barat, "Gibbs sampling methods for Pitman-Yor mixture models," 2014, Research report. [Online]. Available: <https://hal.science/hal-00740770>
- [41] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov Chains*, ser. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2018.
- [42] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*. Springer Science & Business Media, 2013, vol. 293.