Covid19 Reproduction Number: Credibility Intervals by Blockwise Proximal Monte Carlo Samplers

Gersende Fort, Barbara Pascal, Patrice Abry, Nelly Pustelnik

Covid19 Reproduction Number: Credibility Intervals by Blockwise Proximal Monte Carlo Samplers

Gersende Fort, CNRS, Institut de Mathématiques de Toulouse, Toulouse, France
Barbara Pascal, Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
Patrice Abry, CNRS, ENS de Lyon, Laboratoire de Physique, Lyon, France
Nelly Pustelnik, CNRS, ENS de Lyon, Laboratoire de Physique, Lyon, France .

Abstract

Monitoring the Covid19 pandemic constitutes a critical societal stake that received considerable research efforts. The intensity of the pandemic on a given territory is efficiently measured by the reproduction number, quantifying the rate of growth of daily new infections. Recently, estimates for the time evolution of the reproduction number were produced using an inverse problem formulation with a nonsmooth functional minimization. While it was designed to be robust to the limited quality of the Covid19 data (outliers, missing counts), the procedure lacks the ability to output credibility interval based estimates. This remains a severe limitation for practical use in actual pandemic monitoring by epidemiologists that the present work aims to overcome by use of Monte Carlo sampling. After interpretation of the functional into a Bayesian framework, several sampling schemes are tailored to adjust the nonsmooth nature of the resulting posterior distribution. The originality of the devised algorithms stems from combining a Langevin Monte Carlo sampling scheme with Proximal operators. Performance of the new algorithms in producing relevant credibility intervals for the reproduction number estimates and denoised counts are compared. Assessment is conducted on real daily new infection counts made available by the Johns Hopkins University. The interest of the devised monitoring tools are illustrated on Covid19 data from several different countries.

Keywords. Markov chain Monte Carlo sampling, nonsmooth convex optimization, Bayesian inverse problems, credibility interval, Covid19, reproduction number.

I. Introduction

Context. The Covid19 pandemic is causing unprecedented health, social, and economic crises. This triggered massive research efforts to design efficient procedures aiming to assess the intensity of the pandemic, a prerequisite to develop efficient sanitary policies [23]. Several indices are commonly used to measure the strength of a pandemic, such as, e.g., the reproduction number of interest here. However, often, the value of the index alone is not sufficient and credibility intervals of these indices constitute valuable information for the decision makers, notably in periods of rapid pandemic evolution or of changes in trends, an issue not always addressed in pandemic monitoring and at the heart of the present work.

Related works. Pandemic monitoring can be conducted with numerous tools from different scientific fields, (cf. [4] for a review), amongst which *compartmental models*, such as the founding *Susceptible-Infectious-Recovered* scheme. Within pandemic period, when data are scare and of limited quality, the reproduction number, R_t, is often used by epidemiologists, as an efficient practical proxy for the pandemic intensity: it measures the number of second infections caused by one primary infection (cf. e.g., [15, 16, 34, 50, 51]). It thus plays a key role in the pandemic evolution assessment: The number of new infections today, Z_t, depends on R_t and on an average of the new infection counts on previous days $\{\ldots, \mathsf{Z}_{t-3}, \mathsf{Z}_{t-2}, \mathsf{Z}_{t-1}\}$, weighted by the so-called *serial interval function* $\Phi(\cdot)$, that quantifies the distribution of the random delays between the onsets of symptoms in a primary and secondary cases [15, 29, 34, 49]. It has recently been shown that, within pandemic, reliable estimates for the temporal estimation of R_t can be obtained from an inverse problem formulation resulting in a nonsmooth convex optimization problem [2, 37]. The functional to minimize is built from combining the pandemic model in [15], with time regularity constraints. While the procedure was engineered to produce realistic estimations of the temporal evolution of the reproduction number that are robust to the limited quality of the Covid19 pandemic data (severely corrupted with outliers, missing or negative counts and pseudo-seasonalities), it does not however provide credibility intervals, a critical issue towards its practical and actual use by epidemiologists that we aim to address in the present work.

Goals, contributions and outline. The overall goal of the present work is to devise Monte Carlo sampling strategies to perform the estimation by means of credibility intervals of the pandemic reproduction number and of denoised infection counts. To that end, Section II details the proposed Bayesian model used to embed into a stochastic framework the epidemiological model in [15] and its robust extension to data corruption [37]. Its originality stems from using non-differentiable priors to ensure robustness to data corruption. The uniqueness of the maximum a posteriori is thoroughly studied. Further, Section III devises original sampling schemes tailored to handle the non-differentiability of the target distribution: we propose two blockwise

Gersende Fort is with CNRS, Institut de Mathématiques de Toulouse, Toulouse, France (e-mail: gersende.fort@math.univ-toulouse.fr). Work partly supported by the *Fondation Simone et Cino Del Duca, Institut de France*.
Barbara Pascal is with Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France (e-mail: barbara.pascal@univ-lille.fr).
Patrice Abry and Nelly Pustelnik are with CNRS, ENS de Lyon, Laboratoire de Physique, Lyon, France (e-mail: firstname.lastname@ens-lyon.fr).

Proximal-Gradient based extensions of the Langevin Metropolis algorithms: `PGdec` and `PGdual`. We establish their ergodicity, and carry out a comparative study. Using real Covid19 data, made available at the Johns Hopkins University repository and described in Section IV, the performance of up to twelve variations of the sampling strategies are assessed and compared, using well-thought indices quantifying their efficiency (cf. Section V). Finally, in Section VI, the relevance of the proposed blockwise Proximal-Gradient samplers is illustrated for several different countries representative of the evolution of the pandemic across the world, for a 5-week recent period. Daily updates of these credibility interval estimates as well as Matlab routines for their calculations are available on the authors web pages.

Notations. Vectors are column-vectors. For $p \leq q$, the vector $x_{p:q}$ concatenates the scalars or vectors $\{x_i, i = p, \ldots, q\}$. For a matrix A, $\mathsf{A}^\top$ (resp. $\det(\mathsf{A})$ and A^{-1}) denotes the transpose of A (resp. the determinant and the inverse of A). We set $\mathsf{A}^{-\top} := (\mathsf{A}^\top)^{-1} = (\mathsf{A}^{-1})^\top$. I_p is the $p \times p$ identity matrix, and $0_{p\times q}$ is the $p \times q$ null matrix. For a vector $x \in \mathbb{R}^p$, $\|x\|_1$ is the L^1-norm and $\|x\|$ is the L^2-norm. Finally, $\mathcal{N}_r(\mu, \mathsf{C})$ denotes the $\mathbb{R}^r$-valued Gaussian distribution with expectation μ and covariance matrix C. For some $\gamma > 0$, the proximity operator of a proper, convex, lower semi-continuous function f from $\mathbb{R}^d$ to $]-\infty, +\infty]$ is defined as

$$(\forall x \in \mathbb{R}^d) \quad \mathrm{prox}_{\gamma f}(x) := \arg\min_{y\in\mathbb{R}^d} \gamma f(y) + \frac{1}{2}\|y - x\|^2.$$

II. COVID19 PANDEMIC BAYESIAN MODEL

A. Pandemic model

The present work makes use of a pandemic model devised by epidemiologists in [15] that focuses on a main pandemic index: the reproduction number $\mathbf{R}$, to be estimated from daily new infection counts $\mathbf{Z}$. Elaborating on [15], it was further proposed in [37] to account for the limited quality of the intra-pandemic Covid19 data - highly corrupted by irrelevant, missing and mis-reported counts or by pseudo-seasonal effects - by means of additional *outliers* $\mathbf{O}$, also unknown and to be estimated. The goal of the present work is thus to estimate, from a vector of T observed new daily infection counts $\mathbf{Z} := (\mathsf{Z}_1, \ldots, \mathsf{Z}_T)^\top \in \mathbb{N}^T$, the vector of unknowns

$$(\mathbf{R}, \mathbf{O}) := \big((\mathsf{R}_1, \ldots, \mathsf{R}_T)^\top, (\mathsf{O}_1, \ldots, \mathsf{O}_T)^\top\big) \in (\mathbb{R}_+)^T \times \mathbb{R}^T$$

gathering reproduction numbers R_t and outliers O_t.

B. Bayesian model

Estimation entails the recourse to a Bayesian formulation of this pandemic model, where the unknown parameters

$$\boldsymbol{\theta} := (\mathbf{R}, \mathbf{O}),$$

and the observed data $\mathbf{Z}$ are realizations of random vectors Θ and $\mathcal{Z}$, whose distributions need to be specified. We define the a posteriori distribution (which is proportional to the joint distribution by the Bayes theorem), by its density with respect to the Lebesgue measure on $(\mathbb{R}_+)^T \times \mathbb{R}^T$ given by

$$\pi(\boldsymbol{\theta}|\mathbf{Z}) \propto \prod_{t=1}^{T} P(Z_t|\mathbf{Z}_{1:t-1}, \mathsf{R}_t, \mathsf{O}_t) P(\mathsf{R}_t, \mathsf{O}_t|\mathbf{R}_{1:t-1}, \mathbf{O}_{1:t-1})$$

involving the product of the likelihood and of the prior over $(\mathsf{R}_t, \mathsf{O}_t)$, specified below.

Likelihood. The likelihood is modeled as follows: the rate at which a person infected at time $t-u$, generates new infections at time t is $\mathsf{R}_t\Phi_u$ where $\Phi := (\Phi_u)_{1\leq u\leq \tau_\phi}$ denotes the serial interval function, describing the average infectiousness profile after infection [15, 29, 49]. Φ is assumed known and, following [27, 41], is classically modeled as a Gamma distribution truncated over $\tau_\phi = 26$ days with mean and standard deviation of 6.6 and 3.5 days. Then, following [15, 37], the pandemic diffusion is modeled as a Poisson distribution: The conditional distribution of Z_t given the past $\mathbf{Z}_{t-\tau_\phi:t-1}$ and $(\mathsf{R}_t, \mathsf{O}_t)$ reads

$$P(\mathsf{Z}_t|\mathbf{Z}_{t-\tau_\phi:t-1}, \mathsf{R}_t, \mathsf{O}_t) := \frac{\big(p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1})\big)^{\mathsf{Z}_t} \exp(-p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1}))}{\mathsf{Z}_t!} \tag{1}$$

where the intensity $p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1})$ has to be nonnegative; by convention, a Poisson distribution with null intensity is the Dirac mass at zero. The likelihood of the observations $\mathbf{Z}$ is hence defined for any $\boldsymbol{\theta}$ in the measurable set $\mathcal{D}_\mathbf{Z} \subseteq (\mathbb{R}_+)^T \times \mathbb{R}^T$ given by

$$\mathcal{D}_\mathbf{Z} := \{\boldsymbol{\theta} : p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1}) > 0 \text{ for } t \text{ s.t. } \mathsf{Z}_t > 0\} \cup \{\boldsymbol{\theta} : p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1}) \geq 0 \text{ for } t \text{ s.t. } \mathsf{Z}_t = 0\}. \tag{2}$$

A key element of that modeling is that the intensity of the Poisson distribution varies along time as:

$$p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1}) := \mathsf{R}_t \sum_{u=1}^{\tau_\phi} \Phi_u \mathsf{Z}_{t-u} + \mathsf{O}_t. \tag{3}$$

By convention, all distributions are conditionally to initial values $(\mathsf{Z}_{1-\tau_\phi},\dots,\mathsf{Z}_0)$, omitted in notations.

Prior distribution. Regarding the prior, it is assumed that $\mathbf{R}$ and $\mathbf{O}$ are mutually independent. Further, on one hand, the outliers $\mathbf{O}$ are assumed independent and distributed as a Laplace distribution (with parameter $\lambda_{\mathsf{O}}>0$) as commonly encountered in the literature (see, e.g., [22, 33, 36]). The decay of the Laplace distribution in the tails favors some large values among many small ones. This yields

$$P(\mathsf{O}_t) := \frac{\lambda_0}{2}\exp(-\lambda_{\mathsf{O}}|\mathsf{O}_t|). \tag{4}$$

On the other hand, to model smooth piecewise linear time evolutions for R, or equivalently a sparse set of components where the discrete second order derivative in time of R is non zero, $\mathbf{R}$ is assumed distributed as a Laplace AR(2) process: for every $t>2$,

$$P(\mathsf{R}_t|\mathbf{R}_{1:t-1}) := \frac{\lambda_{\mathsf{R}}}{2\sqrt{6}}\exp\left(-\frac{\lambda_{\mathsf{R}}}{\sqrt{6}}|\mathsf{R}_t-2\mathsf{R}_{t-1}+\mathsf{R}_{t-2}|\right), \tag{5}$$

where $\lambda_{\mathsf{R}}>0$. $\lambda_{\mathsf{R}},\lambda_{\mathsf{O}}$ are (fixed) positive *regularization hyperparameters*, balancing the strengths of the different penalizations against the likelihood term. This yields the a priori distribution:

$$\prod_{t=1}^{T} P(\mathsf{R}_t,\mathsf{O}_t|\boldsymbol{\theta}_{1:t-1}) = \prod_{t=3}^{T} P(\mathsf{R}_t|\mathbf{R}_{t-2:t-1})\prod_{t=1}^{T} P(\mathsf{O}_t).$$

Posterior distribution. Combining the likelihood (1) and the priors (4) and (5) leads to the a posteriori density with respect to the Lebesgue measure on $(\mathbb{R}_+)^T\times\mathbb{R}^T$

$$\boldsymbol{\theta}\mapsto\pi(\boldsymbol{\theta}|\mathbf{Z}) \propto \exp\left(-f_{\mathbf{Z}}(\boldsymbol{\theta})-g(\boldsymbol{\theta})\right)1_{\mathcal{D}_{\mathbf{Z}}}(\boldsymbol{\theta}), \tag{6}$$

where 1_A denotes the $\{0,1\}$-valued indicator function of the set A and

$$\begin{cases} f_{\mathbf{Z}}(\boldsymbol{\theta}) := \sum_t d_{\mathsf{KL}}(\mathsf{Z}_t|p(\mathsf{R}_t,\mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1})),\\ g(\boldsymbol{\theta}) := \lambda_{\mathsf{R}}\|\mathsf{D}_2\mathbf{R}\|_1+\lambda_{\mathsf{O}}\|\mathbf{O}\|_1\,; \end{cases} \tag{7}$$

d_{KL} denotes the Kullback-Leibler divergence, related to the log-likelihood of a Poisson process, whose definition is, for some $z\in\mathbb{N}$,

$$(\forall p\in\mathbb{R})\;\; d_{\mathsf{KL}}(z|p) := \begin{cases} z\ln\frac{z}{p}+p-z & \text{if } z>0,\ p>0,\\ p & \text{if } z=0,\ p\geq 0,\\ \infty & \text{otherwise,}\end{cases} \tag{8}$$

and D_2 is the discrete-time second order derivative $(T-2)\times T$ matrix:

$$\mathsf{D}_2 := \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0\\ 0 & 1 & -2 & 1 & 0 & \dots & 0\\ \dots & & & & & & \dots\\ 0 & \dots & & & 1 & -2 & 1\end{bmatrix}. \tag{9}$$

C. Bayesian estimators

In the Bayesian approach to Decision Theory, the maximum, the median, and the expectation of the a posteriori distribution, are Bayes estimators $\widehat{\boldsymbol{\theta}}$ associated to a loss function ℓ

$$\widehat{\boldsymbol{\theta}}(\mathbf{Z}) := \mathrm{Argmin}_{\boldsymbol{\tau}\in\mathcal{D}_{\mathbf{Z}}}\int_{\mathcal{D}_{\mathbf{Z}}}\ell(\boldsymbol{\tau},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Z})\mathrm{d}\boldsymbol{\theta};$$

ℓ is, respectively, the $0-1$ loss, the L^1-norm and the squared L^2-norm (see e.g. [42, Sections 2.3. and 2.5.]). Computing the Maximum a Posteriori (MAP) fits the minimization problem proposed in [37] for the reconstruction of $\boldsymbol{\theta}$:

$$\underset{\mathbf{R},\mathbf{O}}{\text{minimize}}\sum_{t=1}^{T} d_{\mathsf{KL}}(\mathsf{Z}_t|p(\mathsf{R}_t,\mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1}))+\lambda_{\mathsf{R}}\|\mathsf{D}_2\mathbf{R}\|_1+\lambda_{\mathsf{O}}\|\mathbf{O}\|_1\,. \tag{10}$$

The optimization problem is a non-smooth convex minimization problem encapsulating both the transmission process, favoring piecewise linear behavior of the reproduction number along time and sparsity of the outliers. The minimization is performed with the Chambolle-Pock primal-dual algorithm allowing to handle both the non-differentiability and linear operators [7, 10]. Properties of the MAP are established in Proposition 1.

Proposition 1. *If there are at least two positive averaged counts $\sum_{u=1}^{\tau_\phi}\Phi_u\mathsf{Z}_{t_\star-u}$, $\sum_{u=1}^{\tau_\phi}\Phi_u\mathsf{Z}_{t_{\star\star}-u}$, and one positive count $\mathsf{Z}_\tau>0$, a MAP exists. If $\boldsymbol{\theta}^\star=(\mathbf{R}^\star,\mathbf{O}^\star)$ and $\boldsymbol{\theta}^{\star\star}=(\mathbf{R}^{\star\star},\mathbf{O}^{\star\star})$ are two MAP then for any $t\in\{1,\dots,T\}$: $p(\mathsf{R}_t^\star,\mathsf{O}_t^\star|\mathbf{Z}_{t-\tau_\phi:t-1}) = p(\mathsf{R}_t^{\star\star},\mathsf{O}_t^{\star\star}|\mathbf{Z}_{t-\tau_\phi:t-1})$, $\mathsf{O}_t^\star\,\mathsf{O}_t^{\star\star}\geq 0$ and $(\mathsf{D}_2\mathbf{R}^\star)_t\,(\mathsf{D}_2\mathbf{R}^{\star\star})_t\geq 0$.*

Proof. The first statement is adapted from [37]; the second statement is established in [37]. The sign conditions result from a first order expansion of the L^1-norm. For a detailed proof, see section VIII in the supplementary material. □

Proposition 1 implies that the MAP is either unique, or that there are uncountably many MAP. In addition, it shows that $f_{\mathbf{Z}}$ and g are constant over the set of the minimizers. Thus, following the same lines as in [3], a sufficient condition for the uniqueness of the MAP is derived (see section VIII in supplementary material).

The expression of the a posteriori distribution (6) is so complex that the distribution is known only up to a normalizing constant. Consequently, the computation of most statistics of $\pi(\cdot|\mathbf{Z})$ relies on Monte Carlo samplers, in order to produce points $\{\boldsymbol{\theta}^n, n \geq 0\}$ in $\mathcal{D}_{\mathbf{Z}}$ *approximating* π (see e.g. [17, section 2.3]): for example, the estimation of the median and more generally of quantiles of $\pi(\cdot|\mathbf{Z})$ can rely on the order statistics of the points, and the mean a posteriori can be approximated by the Monte Carlo sum $N^{-1}\sum_{n=1}^{N}\boldsymbol{\theta}^n$.

III. Blockwise Proximal-Gradient Monte Carlo samplers

The aim is now to devise Monte Carlo sampling strategies for the *posterior* distribution π defined in (6)-(7). However, this section will address a broader class of densities, defined on $\mathbb{R}^d$ with respect to the Lebesgue measure, and expressed as $\pi(\boldsymbol{\theta}) \propto \exp(-F(\boldsymbol{\theta}))1_{\mathcal{D}}(\boldsymbol{\theta})$ where $F := f + g$ and f, g, and $\mathcal{D}$ satisfy the smoothness and blockwise structure assumptions A1 and A2 defined below.

A. Blockwise structure

A 1. *f and g are finite on $\mathcal{D} \subseteq \mathbb{R}^d$ and f is continuously differentiable on the interior of $\mathcal{D}$.*

Additionally, g has a blockwise structure that we aim to use in the design of the proposed samplers. This blockwise structure stems both from the decomposition of $\boldsymbol{\theta}$ into J blocks $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J) \in \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_J}$ and from the sum of several functions of $\boldsymbol{\theta}_j$ possibly combined with a linear operator.

A 2. *For $j \in \{1, \ldots, J\}$, $i \in \{1, \ldots, I_j\}$, there exist matrices $\mathsf{A}_{i,j} \in \mathbb{R}^{c_{i,j} \times d_j}$, and proper, convex, lower semi-continuous functions $g_{i,j} \colon \mathbb{R}^{c_{i,j}} \to]-\infty, +\infty]$ such that $\sum_{j=1}^{J} d_j = d$ and*

$$\forall \boldsymbol{\theta} := (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_J^\top)^\top, \qquad g(\boldsymbol{\theta}) := \sum_{j=1}^{J}\sum_{i=1}^{I_j} g_{i,j}(\mathsf{A}_{i,j}\boldsymbol{\theta}_j) \; .$$

In addition, the proximity operator of $g_{i,j}$ has a closed form expression.

In Bayesian inverse problems, f may stand for the data fidelity term and the non-smooth part g stands for many penalty terms acting on blocks of the parameter $\boldsymbol{\theta}$. Different splittings of the prior defined by (7) fits Assumption A2.

Example 2. *The prior g given by* (7) *satisfies A2:*

$$\lambda_R\|\mathsf{D}_2\mathbf{R}\|_1 + \lambda_O\|\mathbf{O}\|_1 = g_{1,1}(\mathsf{A}_{1,1}\mathbf{R}) + g_{1,2}(\mathsf{A}_{1,2}\mathbf{O}),$$

where $\boldsymbol{\theta}_1 := \mathbf{R}$, $\boldsymbol{\theta}_2 := \mathbf{O}$, $\mathsf{A}_{1,1} := \mathsf{D}_2$, $\mathsf{A}_{1,2} = \mathrm{I}_T$, $g_{1,1} := \lambda_R\|\cdot\|_1$, and $g_{1,2} := \lambda_O\|\cdot\|_1$.

Example 3. *The prior g given by* (7) *satisfies A2:*

$$\lambda_R\|\mathsf{D}_2\mathbf{R}\|_1 + \lambda_O\|\mathbf{O}\|_1 = \sum_{i=1}^{3} g_{i,1}(\mathsf{A}_{i,1}\mathbf{R}) + g_{1,2}(\mathsf{A}_{1,2}\mathbf{O}),$$

where $\boldsymbol{\theta}_1 := \mathbf{R}$ and $\boldsymbol{\theta}_2 := \mathbf{O}$. $\mathsf{A}_{i,1}$ collects the rows $i, i+3, i+6, \ldots,$ of the matrix D_2, $\mathsf{A}_{1,2} := \mathrm{I}_T$, $g_{i,1} := \lambda_R\|\cdot\|_1$ and $g_{1,2} := \lambda_O\|\cdot\|_1$.

The second example follows block splitting strategies described in [38, 39].

B. Proximal algorithms and Metropolis-Hastings algorithms

The design of an optimization strategy to minimize F on $\mathcal{D}$ and the design of an algorithm to reach the target distribution π both relies on the activation of an operator $\boldsymbol{\mu} \colon \mathbb{R}^d \to \mathbb{R}^d$. To be more specific, when minimizing F, we aim to design a sequence of the form:

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\mu}(\boldsymbol{\theta}^n) \tag{11}$$

where $\boldsymbol{\mu}$ is an operator build from F and $\mathcal{D}$ in such a way that the sequence $(\boldsymbol{\theta}^n)_{n\in\mathbb{N}}$ converges to a minimizer of F (cf. [8, 11, 14] for an exhaustive list of algorithmic schemes). When building a Metropolis-Hastings algorithm (say with Gaussian proposal), a new point is proposed as:

$$\boldsymbol{\theta}^{n+1/2} = \boldsymbol{\mu}(\boldsymbol{\theta}^n) + \xi^{n+1} \quad \text{where} \quad \xi^{n+1} \sim \mathcal{N}_d(0_d, \mathsf{C}) \tag{12}$$

where $\mathsf{C} \in \mathbb{R}^{d\times d}$ is a positive definite matrix. The Langevin dynamics is recovered in the specific case where F is smooth with $\boldsymbol{\mu}(\boldsymbol{\theta}) := \boldsymbol{\theta} - \gamma\nabla F(\boldsymbol{\theta})$ being a gradient ascent over $\ln\pi$ with step size $\gamma > 0$ and $\mathsf{C} := 2\gamma\,\mathrm{I}_d$. Scaled Langevin samplers are also popular: given a $d \times d$ matrix Γ, set

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \boldsymbol{\theta} - \gamma\Gamma\Gamma^\top\nabla F(\boldsymbol{\theta}), \qquad \mathsf{C} := 2\gamma\,\Gamma\Gamma^\top; \tag{13}$$

they are inherited from the so-called *tempered Langevin diffusions* [28] (see also [46] for a pioneering work on its use in the Markov chain Monte Carlo literature). Either this proposed point is the new point $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+1/2}$ (thus yielding the *Langevin Monte Carlo* algorithm [35]; see also [19, 21]) or there is an acceptance-rejection Metropolis mechanism (thus yielding the *Metropolis Adjusted Langevin Algorithm* (MALA) [45]). The general Metropolis-Hastings procedure with Gaussian proposal, is summarized in Algorithm 1 where we denote by $q(\boldsymbol{\theta}, \boldsymbol{\tau})$ the density of the distribution $\mathcal{N}_d(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathsf{C})$ evaluated at $\boldsymbol{\tau} \in \mathbb{R}^d$:

$$\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{\tau}) := \frac{\exp\left(-0.5\,(\boldsymbol{\tau} - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top\,\mathsf{C}^{-1}(\boldsymbol{\tau} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right)}{\sqrt{2\pi}^d\sqrt{\det(\mathsf{C})}}.$$

The constraint $\boldsymbol{\theta} \in \mathcal{D}$ is managed by the acceptance-rejection step (since $\pi(\boldsymbol{\theta}^{n+1/2}) = 0$ when $\boldsymbol{\theta}^{n+1/2} \notin \mathcal{D}$) but not necessarily in the proposal mechanism. The MALA algorithms drift the proposed moves towards areas of high probability for

Algorithm 1: Metropolis-Hastings with Gaussian proposals

Data: a positive definite matrix C; a step-size $\gamma > 0$; a positive integer $N_{\max}$; $\boldsymbol{\theta}^0 \in \mathcal{D}$
Result: A $\mathcal{D}$-valued sequence $\{\boldsymbol{\theta}^n, n \in \{0, \dots, N_{\max}\}\}$

for $n = 0, \dots, N_{\max} - 1$ **do**
- Sample $\xi^{n+1} \sim \mathcal{N}(0, \mathsf{C})$;
- Set $\boldsymbol{\theta}^{n+\frac{1}{2}} = \boldsymbol{\mu}(\boldsymbol{\theta}^n) + \xi^{n+1}$;
- Set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+\frac{1}{2}}$ with probability

$$1 \wedge \frac{\pi(\boldsymbol{\theta}^{n+\frac{1}{2}})}{\pi(\boldsymbol{\theta}^n)}\frac{q(\boldsymbol{\theta}^{n+\frac{1}{2}}, \boldsymbol{\theta}^n)}{q(\boldsymbol{\theta}^n, \boldsymbol{\theta}^{n+\frac{1}{2}})} \tag{14}$$

  and $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$ otherwise.

the distribution π, using first order information on π. Building on this idea, many strategies were proposed in the literature in the setting defined by A1: $\boldsymbol{\mu}$ can either be a gradient step when F is smooth, or a proximal step (i.e., $\boldsymbol{\mu} = \mathrm{prox}_{\gamma F}$ also referred as an implicit subgradient descent step), or a Moreau-Yosida envelope gradient step (i.e. $\boldsymbol{\mu} = \mathrm{I} - \gamma\nabla({}^\gamma F)$ where the Moreau envelope of a function F with parameter $\gamma > 0$ is defined as ${}^\gamma F := \inf_y \gamma F(y) + \frac{1}{2}\|\cdot - y\|^2$). In [18] explicit subgradient steps possibly combined with a proximal step are used. In [13], $\boldsymbol{\mu}$ relies on a Gaussian smoothing of convex functions with Hölder-continuous sub-gradients; this method applies under regularity conditions and convexity assumptions on g which are not implied by A1-A2. In [20], the authors adds a Moreau-Yosida envelope term and a gradient term. In [30], $\boldsymbol{\mu}$ compose a Moreau-Yosida envelope of $g(\mathsf{A}\cdot)$ and a gradient step. Let us cite [6, 47, 48] who also use proximal operators in order to define trans-dimensional Monte Carlo samplers – an objective which is out of the scope defined by A1.

However, in the optimization context (11) with a non-smooth objective function F, proximal-based strategies are often preferred to explicit subgradient steps as the convergence is insured with a fixed step size; explicit subgradient descent steps require decreasing step sizes, far from being numerically efficient. When $F = f + g$ where f and g satisfy Assumption A1, deriving a proximal activation $\boldsymbol{\mu}$ is often a tedious task as no closed form expression for prox_{f+g} exists in a general framework [40, 52]. The standard solution consists in a proximal-gradient activation: $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathrm{prox}_{\gamma g}(\boldsymbol{\theta} - \gamma\nabla f(\boldsymbol{\theta}))$. When dealing with a blockwise structure for g as in A2, the choice of $\boldsymbol{\mu}$ has to manage both the additive structure of g and the combination of the $g_{i,j}$'s with a linear operator. Unfortunately, the proximity operator has a closed form expression in very limited cases recalled below in Lemma 4.

Lemma 4. *Let* $\mathsf{A} \in \mathbb{R}^{c\times d}$.

1) *Let* $g := \frac{1}{2}\|\mathsf{A}\cdot - z\|^2$ *with* $z \in \mathbb{R}^c$. *For every* $\gamma > 0$,

$$\mathrm{prox}_{\gamma g(\mathsf{A}\cdot)} = (\gamma\mathsf{A}^\top\mathsf{A} + \mathrm{I}_d)^{-1}(\cdot + \gamma A^\top z).$$

2) *Let* g *be a proper lower semi-continuous convex function. Let* $\mathsf{A}\mathsf{A}^\top$ *be invertible. For every* $\gamma > 0$,

$$\mathrm{prox}_{\gamma g(\mathsf{A}\cdot)} = \mathrm{I}_d - \mathsf{A}^\top(\mathsf{A}\mathsf{A}^\top)^{-1}(\mathrm{I}_d - \mathrm{prox}_{\gamma g}^{(\mathsf{A}\mathsf{A}^\top)^{-1}})\mathsf{A}$$

where, for every $x \in \mathbb{R}^d$,

$$\mathrm{prox}_{\gamma g}^{(\mathsf{A}\mathsf{A}^\top)^{-1}}(x) := \operatorname*{Argmin}_y\, g(y) + \frac{1}{2\gamma}(x - y)^\top(\mathsf{A}\mathsf{A}^\top)^{-1}(x - y).$$

3) *Let* $g(\mathsf{A}\cdot) := \sum_{\ell=1}^{c} g_\ell(\boldsymbol{a}_\ell \cdot)$ *where* g_ℓ *is convex, lower semi-continuous, and proper from* $\mathbb{R}$ *to* $]-\infty,+\infty]$*, and* $\boldsymbol{a}_\ell \in \mathbb{R}^d$ *denotes the row* $\#\ell$ *of* A*. Suppose that* $\mathsf{A}\mathsf{A}^\top = \Lambda$*, where* $\Lambda := \mathrm{diag}(\chi_1, \ldots, \chi_c)$ *and* $\chi_\ell > 0$*. Then, for every* $\gamma > 0$*,*

$$(\forall \boldsymbol{\eta} \in \mathbb{R}^d) \quad \mathrm{prox}_{\gamma g \circ \mathsf{A}}(\boldsymbol{\eta}) = \boldsymbol{\eta} - \mathsf{A}^\top \Lambda^{-1}\big(\mathsf{A}\boldsymbol{\eta} - \mathrm{prox}_{\gamma \Lambda g}(\mathsf{A}\boldsymbol{\eta})\big)$$

where for all $\boldsymbol{\zeta} = \boldsymbol{\zeta}_{1:c}$*, we set* $\Lambda g(\boldsymbol{\zeta}) := \sum_{\ell=1}^{c} \chi_\ell g_\ell(\zeta_\ell)$*.*

Proof. 1) See [14]. 2) This result is a direct consequence of [7, Proposition 23.25 (ii)] and [7, Proposition 23.345(ii)-(iii)]. 3) Result extracted from [39] and a direct consequence of 2) for specific choices of A and g. □

C. Proposed Blockwise Proximal Monte Carlo Samplers: `PGdec` *and* `PGdual`

None of the algorithms recalled in the previous section directly applies to the context of Assumptions A1 to A2. We propose two novel Metropolis-Hastings algorithms: the `PGdec` sampler and the `PGdual` sampler, which use the proximal operator to handle the non-smooth part g of $-\ln \pi$, its additive structure and its combination with linear operators.

• **The** `PGdec` **sampler.** Additionally to Assumptions A1 and A2, we further assume:

A3. *Each function* $g_{i,j}(\mathsf{A}_{i,j}\cdot)$ *possesses a proximal operator having a closed form expression.*

A3 assumes that each component $g_{i,j}(\mathsf{A}_{i,j}\cdot)$ has a tractable proximal operator which does not imply that the nonsmooth component g admits a tractable proximal operator.

The *Proximal-Gradient Decomposition sampler* (`PGdec`) is described by algorithm 2. It is a Metropolis-Hastings sampler with Gaussian proposal: conditionally to the current point $\boldsymbol{\theta}^n$, `PGdec` proposes a move to $\boldsymbol{\theta}^{n+1/2} = (\boldsymbol{\theta}_1^{n+1/2}, \cdots, \boldsymbol{\theta}_J^{n+1/2})$ sampling independently the J blocks from Gaussian distributions (see $\boldsymbol{\theta}^{n+1/2}$ in line 5 of algorithm 2); then a Metropolis acceptance-rejection step is applied (see (17)). The originality of our method is the definition of $\boldsymbol{\mu}$: for every $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, I_j\}$,

$$\boldsymbol{\mu}_{i,j}^{\text{PGdec}}(\boldsymbol{\theta}) := \mathrm{prox}_{\gamma_j\, g_{i,j}(\mathsf{A}_{i,j}\cdot)}\left(\boldsymbol{\theta}_j - \gamma_j\, \nabla_j f(\boldsymbol{\theta})\right), \tag{15}$$

where γ_j is a positive step size and ∇_j denotes the differential operator with respect to (w.r.t.) the block $\#j$ of $\boldsymbol{\theta}$. The proposed drift takes benefit of the blockwise separable expression of g and computes at each iteration the proximal operator associated to part of the sum in order to perform the proximal activation with a closed form expression. Conditionally to $\boldsymbol{\theta}^n$, for each block $\#j$, one of the component $\#i_j$ is selected at random in $\{1, \ldots, I_j\}$ (see line 3); then, $\boldsymbol{\theta}_j^{n+1/2}$ is sampled from a $\mathbb{R}^{d_j}$-valued Gaussian distribution with expectation $\boldsymbol{\mu}_{i_j,j}^{\text{PGdec}}(\boldsymbol{\theta}^n)$ and covariance matrix $\mathsf{C}_{i_j,j}$. We denote by $q_{i,j}(\boldsymbol{\theta}, \boldsymbol{\tau})$ the density of the distribution $\mathcal{N}_{d_j}(\boldsymbol{\mu}_{i,j}(\boldsymbol{\theta}), \mathsf{C}_{i,j})$ evaluated at $\boldsymbol{\tau} \in \mathbb{R}^{d_j}$:

$$\boldsymbol{q}_{i,j}(\theta, \boldsymbol{\tau}) := \frac{\exp\left(-0.5\,(\boldsymbol{\tau} - \boldsymbol{\mu}_{i,j}(\boldsymbol{\theta}))^\top\, \mathsf{C}_{i,j}^{-1}(\boldsymbol{\tau} - \boldsymbol{\mu}_{i,j}(\boldsymbol{\theta}))\right)}{\sqrt{2\pi}^{d_j}\sqrt{\det(\mathsf{C}_{i,j})}}.$$

Remark 5. *Let* $j \in \{1, \ldots, J\}$ *and* $i \in \{1, \ldots, I_j\}$*. Given* $\boldsymbol{\theta}^n = (\boldsymbol{\theta}_{1:J}^n) \in \mathbb{R}^d$*,* $\boldsymbol{\mu}_{i,j}^{PGdec}(\boldsymbol{\theta}^n)$ *successively computes a gradient step w.r.t. the smooth function* f *and the variable* $\boldsymbol{\theta}_j$*, and a proximal step with respect to the function* $g_{i,j}(\mathsf{A}_{i,j}\cdot)$*; the step size is* γ_j *for both steps. Hence,* $\boldsymbol{\mu}_{i,j}^{PGdec}(\boldsymbol{\theta}^n)$ *is a Proximal-Gradient (PG) step w.r.t. the function*

$$\boldsymbol{\theta}_j \mapsto f(\boldsymbol{\theta}_1^n, \ldots, \boldsymbol{\theta}_{j-1}^n, \boldsymbol{\theta}_j, \boldsymbol{\theta}_{j+1}^n, \ldots, \boldsymbol{\theta}_J^n) + g_{i,j}(\mathsf{A}_{i,j}\boldsymbol{\theta}_j). \tag{16}$$

Example 6 (Example 2 to follow)**.** *The proximal operator of* $\mathbf{R} \mapsto \lambda_R \|\mathsf{D}_2\mathbf{R}\|_1$ *is not explicit, so that, when decomposing* g *as proposed in Example 2,* `PGdec` *can not be applied to sample* $\pi(\cdot|\mathbf{Z})$*.*

Example 7 (Example 3 to follow)**.** *For all* $i \in \{1, 2, 3\}$*,* $\mathsf{A}_{i,1}\mathsf{A}_{i,1}^\top$ *is the identity matrix so that, by Lemma 4,* $\mathrm{prox}_{\gamma_1\, \lambda_R \|\mathsf{A}_{i,1}\cdot\|_1}$ *is explicit and given by* $(\mathrm{I}_T - \mathsf{A}_{i,1}^\top\mathsf{A}_{i,1}) + \mathsf{A}_{i,1}^\top \mathrm{prox}_{\gamma_1\, \lambda_R \|\cdot\|_1}(\mathsf{A}_{i,1}\cdot)$*. In addition,* $\mathrm{prox}_{\gamma_2\, \lambda_O \|\cdot\|_1}$ *has a closed form expression. Hence, when decomposing* g *as proposed in Example 3,* `PGdec` *can be applied to sample* $\pi(\cdot|\mathbf{Z})$*.*

• **The** `PGdual` **sampler.** The *Proximal-Gradient dual sampler* (`PGdual`) is defined along the same lines as algorithm 2. It is designed for situations when for any i, j, the dimensions of the matrices $\mathsf{A}_{i,j}$ satisfy $c_{i,j} \leq d_j$ and $\mathsf{A}_{i,j}$ can be augmented in an invertible $d_j \times d_j$ matrix – denoted by $\bar{\mathsf{A}}_{i,j}$.

For every $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, I_j\}$, let $\bar{\mathsf{A}}_{i,j}$ be a $d_j \times d_j$ invertible matrix such that for any $\boldsymbol{\theta}_j \in \mathbb{R}^{d_j}$, $(\bar{\mathsf{A}}_{i,j}\boldsymbol{\theta}_j)_{d_j - c_{i,j}+1:d_j} = \mathsf{A}_{i,j}\boldsymbol{\theta}_j$; for $\mathbf{x} = (x_1, \ldots, x_{d_j}) \in \mathbb{R}^{d_j}$, define $\bar{g}_{i,j}(\mathbf{x}) := g_{i,j}(x_{d_j - c_{i,j}+1:d_j})$. `PGdual` uses the drift functions

$$\boldsymbol{\mu}_{i,j}^{\text{PGdual}}(\boldsymbol{\theta}) := \bar{\mathsf{A}}_{i,j}^{-1}\mathrm{prox}_{\gamma_j \bar{g}_{i,j}(\cdot)}\left(\bar{\mathsf{A}}_{i,j}\boldsymbol{\theta}_j - \gamma_j \bar{\mathsf{A}}_{i,j}^{-\top}\nabla_j f(\boldsymbol{\theta})\right). \tag{18}$$

Remark 8. *For every* $j \in \{1, \ldots, J\}$*, select* $i_j \in \{1, \ldots, I_j\}$*, and consider the partial objective function:*

$$\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J) + \sum_j \bar{g}_{i_j,j}(\bar{\mathsf{A}}_{i_j,j}\boldsymbol{\theta}_j). \tag{19}$$

Algorithm 2: Blockwise Metropolis-Hastings samplers

Data: $d_j \times d_j$ positive definite matrices $\mathsf{C}_{i,j}$; $\gamma_j > 0$; a positive integer $N_{\max}$; $\boldsymbol{\theta}^0 \in \mathcal{D}$
Result: A $\mathcal{D}$-valued sequence $\{\boldsymbol{\theta}^n, n \in \{0, \dots, N_{\max}\}\}$

for $n = 0, \dots, N_{\max} - 1$ **do**
  for $j = 1, \dots, J$ **do**
    Sample $i_j \in \{1, \dots, I_j\}$ with probability $1/I_j$;
    Sample $\xi_j^{n+1} \sim \mathcal{N}_{d_j}(0_{d_j \times d_j}, \mathsf{C}_{i_j,j})$;
    Set $\boldsymbol{\theta}_j^{n+\frac{1}{2}} = \boldsymbol{\mu}_{i_j,j}(\boldsymbol{\theta}^n) + \xi_j^{n+1}$;
  Set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+\frac{1}{2}}$ with probability

$$1 \wedge \frac{\pi(\boldsymbol{\theta}^{n+\frac{1}{2}})}{\pi(\boldsymbol{\theta}^n)} \prod_{j=1}^{J} \frac{\boldsymbol{q}_{i_j,j}(\boldsymbol{\theta}^{n+\frac{1}{2}}, \boldsymbol{\theta}_j^n)}{\boldsymbol{q}_{i_j,j}(\boldsymbol{\theta}^n, \boldsymbol{\theta}_j^{n+\frac{1}{2}})} \tag{17}$$

  and $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$ otherwise.

Consider the one-to-one maps $\widetilde{\boldsymbol{\theta}}_j = \bar{\mathsf{A}}_{i_j,j}\boldsymbol{\theta}_j$ for any j, and the application

$$\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{1:J} \mapsto f(\bar{\mathsf{A}}_{i_1,1}^{-1}\widetilde{\boldsymbol{\theta}}_1, \dots, \bar{\mathsf{A}}_{i_J,J}^{-1}\widetilde{\boldsymbol{\theta}}_J) + \sum_j \bar{g}_{i_j,j}(\widetilde{\boldsymbol{\theta}}_j). \tag{20}$$

The PG step w.r.t. $\widetilde{\boldsymbol{\theta}}_j$ reads:

$$\mathrm{prox}_{\gamma_j\, \bar{g}_{i,j}}\left(\widetilde{\boldsymbol{\theta}}_j - \gamma_j\, \bar{\mathsf{A}}_{i_j,j}^{-\top} \nabla_j f(\bar{\mathsf{A}}_{i_1,1}^{-1}\widetilde{\boldsymbol{\theta}}_1, \dots, \bar{\mathsf{A}}_{i_J,J}^{-1}\widetilde{\boldsymbol{\theta}}_J)\right) = \mathrm{prox}_{\gamma_j\, \bar{g}_{i,j}}\left(\bar{\mathsf{A}}_{i_j,j}\boldsymbol{\theta}_j - \gamma_j\, \bar{\mathsf{A}}_{i_j,j}^{-\top} \nabla_j f(\boldsymbol{\theta})\right). \tag{21}$$

Therefore, applying a PG step w.r.t. the dual variable $\widetilde{\boldsymbol{\theta}}_j$ in this dual *space, and going back to the original space by applying $\bar{\mathsf{A}}_{i_j,j}^{-1}$ leads to the drift $\boldsymbol{\mu}_{i_j,j}^{PGdual}$ in* (18)*.*

Example 9 (Example 2 to follow)**.** *Denote by $\overline{\mathsf{D}}_2$ any $T \times T$ invertible augmentation of D_2, obtained by adding two vectors in $\mathbb{R}^T$; for example,*

$$\overline{\mathsf{D}}_2 := \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -2/\sqrt{5} & 1/\sqrt{5} & 0 & \dots & 0 \\ & & \mathsf{D}_2 & & \end{bmatrix}; \tag{22}$$

All the rows of the matrix $\overline{\mathsf{D}}_2$ are of norm 1. Then $(\overline{\mathsf{D}}_2\mathbf{R})_{3:T} = \mathsf{D}_2\mathbf{R}$, and for $\mathbf{R} \in \mathbb{R}^T$, $\bar{g}_{1,1}(\mathbf{R}) = \lambda_R\|\mathbf{R}_{3:T}\|_1$. In addition, $\bar{g}_{1,2} = g_{1,2}$. Hence, when decomposing g as proposed in Example 2, `PGdual` *can be used to sample $\pi(\cdot|\mathbf{Z})$.*

• **Choice of the covariance matrix** C. Based on Remark 8, $\bar{\mathsf{A}}_{i,j}\boldsymbol{\mu}_{i,j}^{\mathrm{PGdual}}(\widetilde{\boldsymbol{\theta}})$ is a PG step in a dual space. Since such a step can be seen as an extension of a gradient step to a nonsmooth function, a natural idea is to mime the MALA proposal and add a Gaussian noise with covariance matrix $2\gamma_j\, \mathrm{I}_{d_j}$ in the dual space. Therefore, in the original space, this yields a covariance matrix $\mathsf{C}_{i,j} := 2\gamma_j \bar{\mathsf{A}}_{i,j}^{-1}\bar{\mathsf{A}}_{i,j}^{-\top}$. In (18), note the preconditioning matrix $\bar{\mathsf{A}}_{i,j}^{-\top}$ before the gradient and the matrix $\bar{\mathsf{A}}_{i,j}^{-1}$ before the proximal operator: there is a parallel between such a choice of $\mathsf{C}_{i,j}$ and the scaled MALA proposal mechanism (13).

An equivalent argumentation can be developed for $\boldsymbol{\mu}^{\mathrm{PGdec}}$.

D. Interpretations of `PGdec` *and* `PGdual`

In the case $\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top = \nu_{i,j}\, \mathrm{I}_{c_i}$ with $\nu_{i,j} > 0$, we have from Lemma 4 3),

$$\boldsymbol{\mu}_{i,j}^{\mathrm{PGdec}}(\boldsymbol{\theta}^n) = \left(\mathrm{I}_{d_j} - \Pi_{i,j}\right) \mathtt{G}_j(\boldsymbol{\theta}^n) + \mathsf{A}_{i,j}^\top \left(\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top\right)^{-1} \mathrm{prox}_{\nu_{i,j}\gamma_j g_{i,j}}\left(\mathsf{A}_{i,j}\mathtt{G}_j(\boldsymbol{\theta}^n)\right), \tag{23}$$

where $\mathtt{G}_j(\boldsymbol{\theta}^n) := \boldsymbol{\theta}_j^n - \gamma_j \nabla_j f(\boldsymbol{\theta}^n)$ is a gradient step, and

$$\Pi_{i,j} := \mathsf{A}_{i,j}^\top \left(\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top\right)^{-1} \mathsf{A}_{i,j}$$

is the orthogonal projection matrix on the range of $\mathsf{A}_{i,j}^\top$. (23) shows that $\boldsymbol{\mu}_{i,j}^{\mathrm{PGdec}}(\boldsymbol{\theta}^n)$ is the sum of two orthogonal terms: the first one is the orthogonal projection of $\mathtt{G}_j(\boldsymbol{\theta}^n)$ on the orthogonal space of the range of $\mathsf{A}_{i,j}^\top$, and the second term is in the range space of $\mathsf{A}_{i,j}^\top$. This second term may be seen as a *proximal-contraction* of $\Pi_{i,j}\mathtt{G}_j(\boldsymbol{\theta}^n)$.

Theorem 10 makes the `PGdual` idea explicit by proposing a matrix $\bar{\mathsf{A}}_{i,j}$ augmenting $\mathsf{A}_{i,j}$, computing the associated drift $\boldsymbol{\mu}_{i,j}^{\mathrm{PGdual}}$ and comparing it to $\boldsymbol{\mu}_{i,j}^{\mathrm{PGdec}}$.

Theorem 10. *Assume A1 and A2. Let $j \in \{1,\dots,J\}$ and $i \in \{1,\dots,I_j\}$. Assume that $c_{i,j} < d_j$ and $\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top$ is invertible. Let $\mathsf{U}_{i,j}$ be a $(d_j - c_{i,j}) \times d_j$ matrix such that $\mathsf{U}_{i,j}\mathsf{A}_{i,j}^\top = 0_{(d_j-c_{i,j})\times c_{i,j}}$ and $\mathsf{U}_{i,j}\mathsf{U}_{i,j}^\top$ is invertible. Then, $\bar{\mathsf{A}}_{i,j} := [\mathsf{U}_{i,j};\mathsf{A}_{i,j}] \in \mathbb{R}^{d_j\times d_j}$ is invertible. For any $\boldsymbol{\theta} \in \mathcal{D}$, $\boldsymbol{\mu}_{i,j}^{PGdual}(\boldsymbol{\theta})$ given by* (18) *is equal to*

$$\mathsf{A}_{i,j}^\top \left(\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top\right)^{-1} \mathrm{prox}_{\gamma_j g_{i,j}}\left(\mathsf{A}_{i,j}\left(\boldsymbol{\theta}_j - \gamma_j \widetilde{\Omega}_{i,j}\nabla_j f(\boldsymbol{\theta})\right)\right) + \left(\mathrm{I}_{d_j} - \Pi_{i,j}\right)\left(\boldsymbol{\theta}_j - \gamma_j \Omega_{i,j}\nabla_j f(\boldsymbol{\theta})\right), \tag{24}$$

where

$$\Omega_{i,j} := \mathsf{U}_{i,j}^\top(\mathsf{U}_{i,j}\mathsf{U}_{i,j}^\top)^{-2}\mathsf{U}_{i,j}, \quad \widetilde{\Omega}_{i,j} := \mathsf{A}_{i,j}^\top(\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top)^{-2}\mathsf{A}_{i,j}.$$

Additionally, when $\mathsf{U}_{i,j}\mathsf{U}_{i,j}^\top = \mathrm{I}_{d_j-c_{i,j}}$, $\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top = \mathrm{I}_{c_{i,j}}$, and under A3 then $\boldsymbol{\mu}_{i,j}^{PGdual} = \boldsymbol{\mu}_{i,j}^{PGdec}$ where $\boldsymbol{\mu}_{i,j}^{PGdec}$ is given by (23).

Proof. The main ingredients are the equalities

$$\bar{\mathsf{A}}_{i,j}^{-1} = \left[\mathsf{U}_{i,j}^\top(\mathsf{U}_{i,j}\mathsf{U}_{i,j}^\top)^{-1} \quad \mathsf{A}_{i,j}^\top(\mathsf{A}_{i,j}\mathsf{A}_{i,j}^\top)^{-1}\right];$$

and $\mathsf{U}_{i,j}^\top(\mathsf{U}_{i,j}\mathsf{U}_{i,j}^\top)^{-1}\mathsf{U}_{i,j} + \Pi_{i,j} = \mathrm{I}_{d_j}$. The proof follows from standard matrix algebra. A detailed proof is given in section IX of the Supplementary material. □

The result remains true when $c_{i,j} = d_j$; in that case, $\bar{\mathsf{A}}_{i,j} = \mathsf{A}_{i,j}$, $\mathrm{I}_{d_j} - \Pi_{i,j} = 0_{d_j\times d_j}$ and $\widetilde{\Omega}_{i,j} = \mathsf{A}_{i,j}^{-1}\mathsf{A}_{i,j}^{-\top}$. Theorem 10 provides sufficient conditions for the drift function $\boldsymbol{\mu}_{i,j}^{\mathrm{PGdec}}$ and the drift function $\boldsymbol{\mu}_{i,j}^{\mathrm{PGdual}}$ to be equal. Observe that the conditions on $\mathsf{U}_{i,j}$ are satisfied as soon as the rows of $\mathsf{U}_{i,j}$ are orthonormal and orthogonal to the rows of $\mathsf{A}_{i,j}$.

Let us derive two strategies for the application of `PGdual` to sample the target density defined by (6)-(7).

Example 11 (Example 2 and Example 9, to follow)**.** *A first strategy is to decompose g as in Example 2; Example 9 provides a possible augmentation of D_2. Another one is proposed by Theorem 10:*

$$\overline{\mathsf{D}}_o := \begin{bmatrix}\mathsf{U}_{1,1}\\ \mathsf{D}_2\end{bmatrix} \in \mathbb{R}^{T\times T} \tag{25}$$

where $\mathsf{U}_{1,1}$ is obtained by making orthogonal the first two rows of $\overline{\mathsf{D}}_2$ and making them orthogonal to the rows of D_2. This strategy acts globally on $\ln\pi(\cdot|\mathbf{Z})$ by proposing, at each iteration, a PG approach for the function $f_{\mathbf{Z}} + \lambda_R\|\mathsf{D}_2\cdot\|_1 + \lambda_O\|\cdot\|_1$. It will be numerically explored in section V.

Example 12 (Example 3 to follow)**.** *A second strategy is to decompose g as in Example 3, and define the matrices $\bar{\mathsf{A}}_{i,j}$ as described in Theorem 10. This strategy defines $\boldsymbol{\mu}^{\mathrm{PGdual}}$ by considering part of $\ln\pi(\cdot|\mathbf{Z})$: at each iteration, having selected $i_1 \in \{1,2,3\}$, it proposes a PG approach for the function $f_{\mathbf{Z}} + \lambda_R\|\mathsf{A}_{i_1,1}\cdot\|_1 + \lambda_O\|\cdot\|_1$ (see Remark 8).*

E. Convergence analysis of `PGdec` *and* `PGdual`.

We prove that both `PGdec` and `PGdual` produce a sequence of points $\{\boldsymbol{\theta}^n, n \geq 0\}$ which is an ergodic Markov chain having π as its unique invariant distribution.

Proposition 13. *Assume A1, A2 and A3. Assume also that π is continuous on $\mathcal{D}$. Then the sequence $\{\boldsymbol{\theta}^n, n \geq 0\}$ given by algorithm 2 applied with $\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathrm{PGdec}}$ is a Markov chain, taking values in $\mathcal{D}$ and with unique invariant distribution π. In addition, for any initial point $\boldsymbol{\theta}^0 \in \mathcal{D}$ and any measurable function h such that $\int |h(\boldsymbol{\theta})|\,\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < \infty$,*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} h(\boldsymbol{\theta}^n) = \int h(\boldsymbol{\theta})\,\pi(\mathrm{d}\boldsymbol{\theta}), \quad \textit{with probability one.}$$

The same result holds when $\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathrm{PGdual}}$.

Proof. We have $\boldsymbol{\theta}^{n+1} \in \mathcal{D}$ since $\pi(\boldsymbol{\theta}^{n+1/2}) = 0$ when $\boldsymbol{\theta}^{n+1/2} \notin \mathcal{D}$. The other properties result from [31, Lemmas 1.1. and 1.2] and [32, Propositions 10.1.1 and 10.4.4 and Theorem 17.0.1]. The Harris recurrence property (required for the law of large numbers to hold for any $\boldsymbol{\theta}^0$) can be proved along the same lines as [44, Theorem 6(v)]. See the detailed proof in section IX of the Supplementary material. □

Remark 14. *A slight adaption of the proof shows that when selecting the indices i_j, the uniform distribution on $\{1,\dots,I_j\}$ can be replaced with a probability distribution $\{\rho_{i,j}(\boldsymbol{\theta}^n), i = 1,\dots,I_j\}$ depending on the current value $\boldsymbol{\theta}^n$, up to a modification of the acceptance-rejection ratio* (17) *(see section IX of the Supplementary material).*

F. `Gibbs PGdec` *and* `Gibbs PGdual` *samplers.*

The additive structure of g in A2, naturally suggests the extension of `PGdec` and `PGdual` to the Gibbs sampler algorithm [26]. Since whatever $\boldsymbol{\theta} = \boldsymbol{\theta}_{1:J} \in \mathbb{R}^d$, exact sampling from the conditional distributions on $\mathbb{R}^{d_j}$

$$\boldsymbol{\tau} \mapsto \pi_j(\boldsymbol{\tau}|\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta}_{1:j-1}, \boldsymbol{\tau}, \boldsymbol{\theta}_{j+1:J}) - \sum_{i=1}^{I_j} g_{i,j}(\mathsf{A}_{i,j}\boldsymbol{\tau}))1_{\mathcal{D}}(\boldsymbol{\theta}_{1:j-1}, \boldsymbol{\tau}, \boldsymbol{\theta}_{j+1:J})$$

for $j = 1, \ldots, J$, is not always explicit, we propose a Metropolis-within-Gibbs strategy [12]. The pseudo-code of the so-called `Gibbs Blockwise Proximal` sampler is given by algorithm 3 in the case of a systematic scan order of the J components (our method easily extends to other scan orders; details are left to the reader): at each iteration $\#(n+1)$, and for each block $\#j$, *(i)* sample at random $i_j \in \{1, \ldots, I_j\}$, *(ii)* sample a candidate from $\mathcal{N}_{d_j}(\boldsymbol{\mu}_{i_j,j}(\boldsymbol{\vartheta}), \mathsf{C}_{i_j,j})$ where $\boldsymbol{\vartheta} := (\boldsymbol{\theta}_{1:j-1}^{n+1}, \boldsymbol{\theta}_{j:J}^{n}) \in \mathbb{R}^d$ is the current value of the chain, and *(iii)* accept/reject this candidate via a Metropolis step targeting the distribution $\pi_j(\cdot|\boldsymbol{\vartheta})$.

Algorithm 3: Blockwise Gibbs sampler

Data: $d_j \times d_j$ positive definite matrices $\mathsf{C}_{i,j}$; $\gamma_j > 0$; a positive integer $N_{\max}$; $\theta^0 \in \mathcal{D}$
Result: A $\mathcal{D}$-valued sequence $\{\boldsymbol{\theta}^n, n \in \{0, \ldots, N_{\max}\}\}$

Set $\boldsymbol{\vartheta} = \boldsymbol{\theta}^0$;
for $n = 0, \ldots, N_{\max} - 1$ **do**
  for $j = 1, \ldots, J$ **do**
    Sample $i_j \in \{1, \ldots, I_j\}$ with probability $1/I_j$;
    Sample $\xi_j^{n+1} \sim \mathcal{N}_{d_j}(0_{d_j \times d_j}, \mathsf{C}_{i_j,j})$;
    Set $\boldsymbol{\theta}_j^{n+\frac{1}{2}} = \boldsymbol{\mu}_{i_j,j}(\boldsymbol{\vartheta}) + \xi_j^{n+1}$;
    Set $\boldsymbol{\theta}_j^{n+1} = \boldsymbol{\theta}_j^{n+\frac{1}{2}}$ with probability

$$1 \wedge \frac{\pi_j(\boldsymbol{\theta}_j^{n+\frac{1}{2}}|\boldsymbol{\vartheta})}{\pi_j(\boldsymbol{\theta}_j^{n}|\boldsymbol{\vartheta})} \frac{q_{i_j,j}((\boldsymbol{\vartheta}_{1:j-1}, \boldsymbol{\theta}_j^{n+\frac{1}{2}}, \boldsymbol{\vartheta}_{j+1:J}), \boldsymbol{\theta}_j^{n})}{q_{i_j,j}(\boldsymbol{\vartheta}, \boldsymbol{\theta}_j^{n+\frac{1}{2}})}$$

    and $\boldsymbol{\theta}_j^{n+1} = \boldsymbol{\theta}_j^{n}$ otherwise ;
    Update $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_{1:j}^{n+1}, \boldsymbol{\theta}_{j+1:J}^{n})$

The `Gibbs PGdec` and the `Gibbs PGdual` samplers correspond to algorithm 3 applied resp. with $\boldsymbol{\mu}_{i,j} = \boldsymbol{\mu}_{i,j}^{\mathrm{PGdec}}$ and $\boldsymbol{\mu}_{i,j} = \boldsymbol{\mu}_{i,j}^{\mathrm{PGdual}}$.

IV. Covid-19 data

To illustrate, assess, and compare the relevance and performance of the Monte Carlo procedures for credibility interval estimation proposed here, use is made of real Covid-19 data made available by the *Johns Hopkins University*[1]. The repository[2] collects, impressively since the outbreak of the pandemic, daily new infections and new death counts from the National Public Health Authorities of 200+ countries and territories of the world. Counts are updated on a daily basis since the early stage of the pandemic (Jan. 1st, 2020) until today. This repository thus provides researchers with a remarkable dataset to analyse the pandemic.

As mentioned in the Introduction section, because of the sanitary crisis context, data made available by most National Public Health Authorities in the world are of very limited quality as they are corrupted by outliers and missing or negative counts. Data quality also varies a lot across countries or even within a given country depending on the phases and periods of the pandemic.

The present work uses the new infection counts only, as the estimation of the space-time evolution of the pandemic reproduction number R is targeted.

A mild and non-informative preprocessing is applied to data by replacing negative counts by a null value.

V. Monte Carlo sampler assessments

The present section aims to assess and compare the performance for several variations of the samplers `PGdec` and `PGdual` introduced in section III, using the real Covid19 data described in section IV.

[1] https://coronavirus.jhu.edu/
[2] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_time_series/

A. Assessment set-up

Performance assessments were conducted on data from several countries and for different periods of interests. For space reasons, they are reported only for the United Kingdom and for a recent time period (Dec. 6th, 2021 to Jan. 9th, 2022), with corresponding $\mathbf{Z}$ in Fig. 2[top right plot].

Following [37], we set $\lambda_{\mathrm{O}} = 0.05$ and $\lambda_{\mathrm{R}} = 3.5 \times \sqrt{6}\,\sigma_{\mathbf{Z}}/4$ with $\sigma_{\mathbf{Z}}$ the standard deviation of $\mathbf{Z}$. Samplers are run with $N_{\max} = 1e7$ iterations, including a *burn-in phase* of $3e6$ iterations. Except for the plots in Fig. 1[first row], reported performance are computed by discarding the points produced during the burn-in phase. The initial values $\mathsf{Z}_{-\tau_\phi+1}, \ldots, \mathsf{Z}_{-1}, \mathsf{Z}_0$ are set to the observed counts. For each algorithm, Performances are computed from averages over 15 independent runs. Initial points consist of random perturbations around the non-informative point $\mathbf{R}^0 := (1, \ldots, 1)^\top \in \mathbb{R}^T$ and $\mathbf{O}^0 := (0, \ldots, 0)^\top \in \mathbb{R}^T$. For all samplers, $\boldsymbol{\theta}$ is seen as a two-block vectors $(\boldsymbol{\theta}_1 = \mathbf{R}, \boldsymbol{\theta}_2 = \mathbf{O})$ (hence $J = 2$).

B. Samplers

`PGdec` **and** `Gibbs PGdec`**.** These two samplers are run by decomposing g as described in Example 3. This yields two different drift functions for the blocks $\mathbf{R}$ and $\mathbf{O}$; see Example 7. For the $\mathbf{R}$-part, the drift functions are defined, for $i = 1, 2, 3$, as:

$$\boldsymbol{\mu}_{i,1}^{\mathrm{PGdec}}(\boldsymbol{\theta}) := \left(\mathsf{I}_T - \mathsf{A}_{i,1}^\top \mathsf{A}_{i,1}\right)\left(\mathbf{R} - \gamma_1 \nabla_1 f_{\mathbf{Z}}(\boldsymbol{\theta})\right) + \mathsf{A}_{i,1}^\top \operatorname{prox}_{\gamma_1 \lambda_{\mathrm{R}} \|\cdot\|_1}\left(\mathsf{A}_{i,1}\mathbf{R} - \gamma_1 \mathsf{A}_{i,1} \nabla_1 f_{\mathbf{Z}}(\boldsymbol{\theta})\right), \tag{26}$$

obtained from (23) with $\mathsf{A}_{i,1}\mathsf{A}_{i,1}^\top$ the identity matrix. For the $\mathbf{O}$-part,

$$\boldsymbol{\mu}_{1,2}^{\mathrm{PGdec}}(\boldsymbol{\theta}) := \operatorname{prox}_{\gamma_2 \lambda_{\mathrm{O}} \|\cdot\|_1}\left(\mathbf{O} - \gamma_2 \nabla_2 f_{\mathbf{Z}}(\boldsymbol{\theta})\right), \tag{27}$$

where ∇_1 (resp. ∇_2) denotes the gradient operator w.r.t $\mathbf{R}$ (resp. $\mathbf{O}$).

`PGdua`**l and** `Gibbs PGdual`**.** These two samplers are run by decomposing g as described in Example 2. Consequently, for the $\mathbf{R}$-part, the drift function is given by

$$\boldsymbol{\mu}_{1,1}^{\mathrm{PGdual}}(\boldsymbol{\theta}) := \overline{\mathsf{D}}^{-1} \operatorname{prox}_{\gamma_1 \lambda_{\mathrm{R}} \|(\cdot)_{3:T}\|_1}\left(\overline{\mathsf{D}}\mathbf{R} - \gamma_1 \overline{\mathsf{D}}^{-\top} \nabla_1 f_{\mathbf{Z}}(\boldsymbol{\theta})\right) \tag{28}$$

where $\overline{\mathsf{D}} := \overline{\mathsf{D}}_2$ (see Example 9, the sampler is referred to as `PGdual Invert (I)`) or $\overline{\mathsf{D}} := \overline{\mathsf{D}}_o$ (see Example 11, the sampler is referred to as `PGdual Ortho (O)`); for the $\mathbf{O}$-part, for both `PGdual I` and `PGdual O`,

$$\boldsymbol{\mu}_{1,2}^{\mathrm{PGdual}}(\boldsymbol{\theta}) := \operatorname{prox}_{\gamma_2 \lambda_{\mathrm{O}} \|\cdot\|_1}\left(\mathbf{O} - \gamma_2 \nabla_2 f_{\mathbf{Z}}(\boldsymbol{\theta})\right) . \tag{29}$$

RW and Gibbs RW. For comparisons, we also run Random Walk-based samplers (`RW`) with Gaussian proposals: `RW` and `Gibbs RW` are defined respectively from algorithm 2 and algorithm 3, with $I_1 = I_2 = 1$ and $\boldsymbol{\mu}_{1,j}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ for $j = 1, 2$.

Covariance matrices. For the covariance matrices $\mathsf{C}_{i,j}$, we choose $\mathsf{C}_{1,2} := 2\gamma_2 \mathsf{I}_T$ for the $\mathbf{O}$-part, where γ_2 is the same step size as in (27) and (29). For the $\mathbf{R}$-part, based on the comment in subsection III-C, we choose $\mathsf{C}_{1,1} := 2\gamma_1 \overline{\mathsf{D}}_2^{-1} \overline{\mathsf{D}}_2^{-\top}$ for `PGdual I` and `Gibbs PGdual I`; and $\mathsf{C}_{1,1} := 2\gamma_1 \overline{\mathsf{D}}_o^{-1} \overline{\mathsf{D}}_o^{-\top}$ for `PGdual O` and `Gibbs PGdual O`. Again, γ_1 is the same step size as in (26) and (28). We also run the `PGdec`-based samplers and the `RW`-based samplers with the same covariance matrices: when $\mathsf{C}_{i,1} := 2\gamma_1 \overline{\mathsf{D}}_2^{-1} \overline{\mathsf{D}}_2^{-\top}$ this yields `PGdec I`, `Gibbs PGdec I`, `RW I` and `Gibbs RW I`; when $\mathsf{C}_{i,1} := 2\gamma_1 \overline{\mathsf{D}}_o^{-1} \overline{\mathsf{D}}_o^{-\top}$, this yields `PGdec O`, `Gibbs PGdec O`, `RW O` and `Gibbs RW O`.

Twelve sampling strategies. The present section will compare twelve different samplers, constructed from the three drift proposition strategies (`PGdec, PGDual, RW`), times two families (`Metropolis-Hastings, Gibbs`), times two choices for the covariance matrix $\mathsf{C}_{i,1}$ (`I, O`).

Step sizes. Different strategies are compared for the definition of the step sizes (γ_1, γ_2). All of them consist in adapting the step sizes during the burn-in phase in such a way that the mean acceptance-rejection rate reaches approximately 0.25 (which is known to be optimal for some `RW` Metropolis-Hastings samplers, [24]). We observed that convergence occurs before $5e5$ iterations, see Fig. 1[row 1, columns 2 and 3] where the proposals are all rejected and the chain does not move from its initial point during the first iterations, when the step size is too large. At the end of the burn-in phase, the step sizes are frozen and no longer adapted. For the Metropolis-Hastings samplers `PGdec`, `PGdual`, and `RW`, γ_1 is adapted and $\gamma_2 := \gamma_1 \sigma_{\mathbf{Z}}^3 / T$. For the Gibbs samplers, the acceptance-rejection steps are specific to each block $\mathbf{R}$ and $\mathbf{O}$. A first consequence is that a move on one block can be accepted while the other one is not; this may yield larger step sizes (γ_1, γ_2) which in turn favor larger moves of the chains. A second consequence is that we use the acceptance-rejection rate for the $\mathbf{R}$-part (resp. for the $\mathbf{O}$-part) to adapt γ_1 (resp. γ_2): the two step sizes have their own efficiency criterion.

C. Performance assessment criteria

The sampler performances are assessed and compared using three different criteria, see Fig. 1.

Distance to the MAP. We compute the normalized distance $\|\mathbf{R}^n - \widehat{\mathbf{R}}_{\mathrm{MAP}}\|/\|\widehat{\mathbf{R}}_{\mathrm{MAP}}\|$ where $\widehat{\mathbf{R}}_{\mathrm{MAP}}$ denotes the MAP estimator, (computed as in [37]): it is displayed vs. the iteration index n, in the burn-in phase (row 1) and after the burn-in phase (row 2). This criterion quantifies the ability of the chains to perform a relevant exploration of the distribution: The chains have to visit the support of $\pi(\cdot|\mathbf{Z})$, they show better ergodicity rates when they are able to rapidly escape from low density regions to move to higher density regions. Paths start from a non-informative initial point, considered as a point in a low density region. This criterion permits to quantify a relevant behavior of the Markov chain when *(i)* it drifts rapidly towards zero during the burn-in phase and *(ii)* it fluctuates in a large neighborhood of zero after the burn-in phase.

Autocorrelation function (ACF). We then compute the ACF for each of the $2T$ components of the vector $\boldsymbol{\theta}$ along the Markov path, for a lag from 1 to $1e5$. On row 3, we report the mean value, over these $2T$ components, of the absolute value of the ACF versus the lag. This criterion quantifies a relevant behavior of the Markov chain when it converges rapidly to zero; it is indeed related to the effective sample size of a Markov chain (see [43]) and to the mixing properties of the chain (see e.g. [32, Theorem 17.4.4 and Section 17.4.3.]). For example, a weaker ACF means that less iterations of the sampler are required to reach a given estimation accuracy.

Gelman-Rubin (GR) statistic. Finally, we compute the GR statistic (see [9, 25]) which quantifies a relevant behavior of the Markov chain when it converges rapidly to one. It measures how the sampler forgets its initial value and provides homogeneous approximations of the target distribution $\pi(\cdot|\mathbf{Z})$ after a given number of iterations. On row 4, we report this statistic versus the iteration index.

D. Performance comparisons.

Covariance matrices. Normalized Distance to MAP indices (Fig. 1[rows 1 & 2]) and GR indices (Fig. 1[row 4]) show that, for all algorithms `RW`, `PGdec`, and `PGdual`, the strategy `O` is more efficient than the strategy `I`, as corresponding indices decay more rapidly (to 0 and 1 respectively) for the formers than for the latters.

Metropolis-Hastings vs. Gibbs samplers. The evolution of the ACF criteria (Fig. 1[row 3]) shows that the Gibbs strategies are globally more efficient than the Metropolis-Hastings ones. Further, GR indices confirm better efficiency of the Markov chains obtained from `Gibbs O` strategies.

Drift functions. The benefit of using functions $\boldsymbol{\mu}$ miming optimization algorithms to drift the proposed points towards the higher probability regions of $\pi(\cdot|\mathbf{Z})$ is clearly quantified in Fig. 1 across all performance indices.

During the burn-in phase, the normalized distance to MAP indices decay toward 0 significantly faster for all `PGdual`-based algorithms, compared to `RW`-based ones. Also, the `PGdual O` algorithms reach an expected plateau early in the burn-in phase, while this is barely the case at the end of the burn-in phase for `RW O` algorithms.

After the burn-in phase (second row), the samplers using optimization-based drift functions $\boldsymbol{\mu}$ show better behavior after reaching the high density regions as they permit a broader and more rapid exploration of the support of the distribution around its maximum, with large amplitude moves from $\widehat{\mathbf{R}}_{\mathrm{MAP}}$, and faster returns to $\widehat{\mathbf{R}}_{\mathrm{MAP}}$. This is notably clearly visible for the `Gibbs PGdual O` strategy.

ACF indices decay more rapidly for the `PGdual` strategies than for the `RW` ones, irrespective of the choice of the covariance matrix or of the Metropolis-Hastings or Gibbs versions. Also, ACF indices for `PGdual` algorithms are less sensitive to the choice of the covariance matrices than the `RW` ones.

Finally, the GR statistic indices (row 4) clearly illustrates that Markov chains produced by `PGdual` samplers have better mixing properties, compared to others, showing hence sensitivity to the choice of the initial point.

Optimal sampler. Combined together, these observations, globally consistent with those stemming from other countries or time periods, lead to the following generic comments.

While yielding essentially equivalent performance, across time periods and countries, the choice of the covariance matrices (algorithms `O` and `I`) significantly outperforms algorithms relying on Identity covariance matrices (not reported here as showing poor performances). This non trivial construction actually stemmed from the thorough and detailed mathematical analysis conducted in subsection III-C. Theorem 10 advocates to perform the augmentation of the $(T-2)\times T$ matrix D_2 into a $T\times T$ invertible matrix $\overline{\mathsf{D}}$, by adding two rows which are orthogonal to the rows of D_2.

The observation that Gibbs samplers show better performances may stem from the fact that they benefit from larger values of the step sizes learned on the fly by the algorithms which favor larger jumps when proposing $\boldsymbol{\theta}^{n+1/2}$ from $\boldsymbol{\theta}^n$ and imply better mixing properties.

Finally, the `PGdual` algorithms show homogeneous performances for example when varying the covariance matrix, and are thus less sensitive to parameter tuning, an important practical feature. Also, the `PGdual` algorithms show systematically better performances than the `PGdec` ones. This may result from the definition of $\boldsymbol{\mu}^{\mathrm{PGdec}}$ which, because of the block-splitting approach (see (26)), uses only partial information on $\pi(\cdot|\mathbf{Z})$ at each iteration.

As an overall conclusion, systematically observed across all studied time periods and countries, the `Gibbs PGdual O` algorithm, devised from the careful analysis of the theoretical properties of the distributions defined by A1-A2, is consistently found to be the most efficient strategy for a relevant assessment of the Covid19 pandemic time evolution.

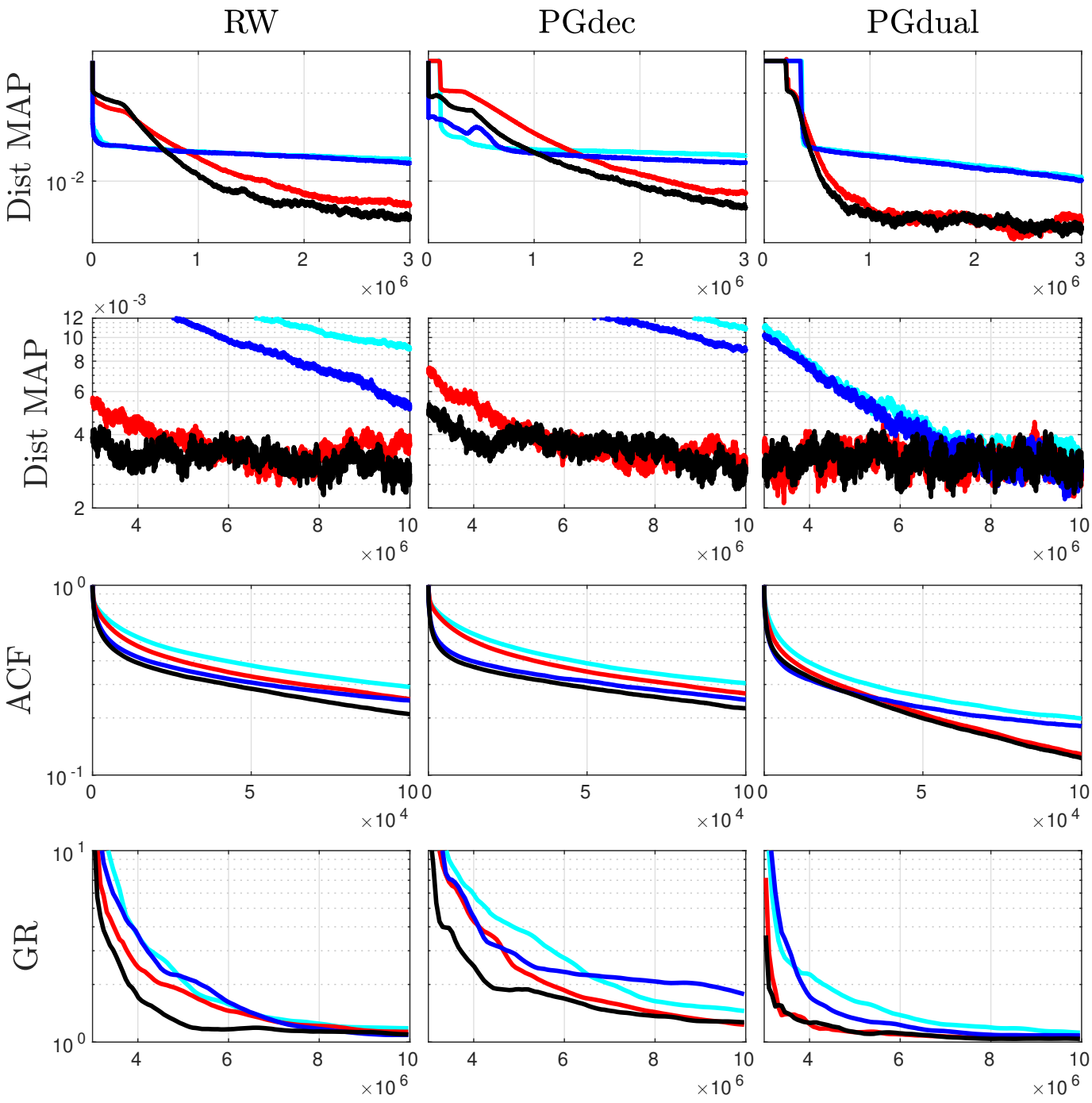

Fig. 1: Displayed: the normalized distance to $\widehat{\mathbf{R}}_{\mathrm{MAP}}$ versus the iteration index, during the burn-in phase (row 1) and after the burn-in phase (row 2); the ACF criterion versus the lag (row 3) and the GR criterion versus the iteration index (row 4); for the `RW`-based samplers (column 1), `PGdec`-based samplers (column 2) and `PGdual`-based samplers (column 3). For each `Algo` in { `RW`, `PGdec`, `PGdual` }: `Algo I` is in cyan, `Algo O` is in red, `Gibbs Algo I` is in blue and `Gibbs Algo O` is in black.

VI. Credibility intervals for R

Goal and set-up. The present section aims to illustrate the relevance of credibility interval-based estimations for the reproduction number R and for the outliers O from real Covid19 data. Because it is of greater interest for epidemiologists, the credibility intervals are translated into credibility intervals for the denoised counts $\mathbf{Z}^{(D)}$ by their simple subtraction to the original counts $\mathbf{Z}$, i.e., intuitively $\mathbf{Z}^{(D)} = \mathbf{Z} - \mathbf{O}$.

Credibility interval-based estimations are reported for `Gibbs PGdual O` sampling scheme only, as Section V established that it achieved the best performance amongst the twelve sampling schemes tested. Estimations are computed for a time period of five weeks ($T = 35$ days), which corresponds to a few typical pandemic time scales, of the order of 7 days, induced by the serial interval function Φ, cf. Section II. This period is set to a recent phase of the pandemic (Dec. 13th, 2021 to Jan. 17th, 2022). Estimations are reported for several countries, arbitrarily chosen as representatives of the pandemic across the world, but conclusions are valid for most countries.

Computation of the credibility intervals. Credibility intervals are computed as follows: For each day $t \in [T_1, T_2]$ in the period of interest $[T_1, T_2]$, the chosen sampling scheme (`Gibbs PGdual O`) outputs 7.10^6 points of a Markov chain approximating the a posteriori distribution of R_t ; Quantiles of that distribution are estimated using the empirical cumulative distribution function ; For a chosen credibility level of $1 - \alpha$, the upper and lower limits of the credibility intervals are defined by the empirical $1 - \alpha/2$ and $\alpha/2$ quantiles. For illustration purposes, $1 - \alpha$ is set here to 95%. The same procedure is applied to produce credibility intervals for the outliers O_t. Finally, credibility intervals for the estimated denoised new infection counts $\mathbf{Z}^{(D)}$ are obtained by subtracting the credibility intervals for O_t to the raw and possibly corrupted count Z_t. Fig. 2 reports, top plots, the daily counts of new infections (black lines) to which are superimposed the 95% credibility interval-based estimations for the denoised counts, $\mathbf{Z}^{(D)}$, (red pipes). Fig. 2 further reports, bottom plots, the 95% credibility interval-based estimations for R.

Relevance of credibility interval-based estimations. Fig. 2, together with the examination of equivalent plots for other countries, yields the following generic conclusions.

The estimated denoised counts show far smoother evolution along time compared to the raw counts, hence providing far more realistic assessments of the intensity of the pandemics. Notably, for most countries, the zero or low counts, associated with week-ends or non-working days, followed by days with over-evaluated counts by compensation, are smoothed out by the outlier estimation procedure, while the values of the counts for the regular (or non corrupted) days are left unchanged. This is the direct benefit of the nonlinear filtering procedure underlying the estimation formulation in (6)-(7), as opposed to traditional denoising procedures performed by classical linear filtering (such as moving average).

For the credibility intervals for R, their sizes range, depending on countries and time periods, from below 1% to above 10%. A careful examination shows that the credibility interval size is mostly driven by data quality: the credibility interval size increases when outliers are detected. Further, for a given country, the size of the credibility intervals varies only mildly along time over a five-week period. Changes in size are often associated with changes in the trends of the estimates of R, or with the occurrence of outliers. These credibility intervals provide hence a relevant assessment not only of the intensity of the pandemics, but also of the confidence that can be granted to this assessment, by providing epidemiologists with a range of likely values of R, rather than a single value. This permits to compare the evolution of the pandemics across several countries on a better scientifically grounded basis.

These credibility interval-based estimations permit a double analysis of the pandemic: They permit retrospectively to evaluate the impacts of sanitary measures on the pandemic evolution. Additionally, the smooth nature of the estimation of R (close to piecewise linear) performs an implicit short term forecast (or a *nowcast*) of the evolution of the pandemic intensity: For instance, for several countries (e.g., France, Mali, Brazil, Singapore,...) the estimate of R is decreasing for the last 5 to 10 days of the studied period, predicting that daily new infections will reach a maximum of the current wave within the coming days and then will start to decrease.

These credibility interval estimates can be complemented with other estimates such as the Maximum, median or Mean a Posteriori (cf. e.g., [1, 5]).

Finally, let us emphasize that these estimates, denoised counts and credibility intervals, are obtained using a single and same set of hyperparameters $\lambda_{\mathrm{R}}/\sigma_{\mathbf{Z}}$ and λ_{O} common to all countries.

VII. Conclusions

The proposed tools perform a relevant credibility interval-based estimation of the Covid19 reproduction number and denoised new infection counts, by combining a Bayesian modeling of the time evolution of the pandemic with Monte Carlo Metropolis sampling strategies. Robustness against the low quality of the Covid19 is achieved by engineering the Bayesian model to impose sparsity in the changes of a smooth time evolution for the reproduction number and in the outlier occurrences, modeling data corruption. This is obtained at the price of the non-differentiability of their a posteriori distribution, thus precluding the use of the classical Metropolis Adjusted Langevin Algorithm to produce credibility intervals. This lead us to propose several Proximal-Gradient Monte Carlo algorithms tailored to the sampling of non-differentiable a posteriori distributions, that also constitute

valid sampling schemes for a much broader range of applications than that of the strict Covid19 pandemic monitoring, e.g., in image processing or more generally, in any Bayesian inverse problems with several non-smooth priors.

Estimation performances were assessed and compared on real Covid19 data, using a set of well-selected indices quantifying the efficiency of these different sampling schemes.

Finally, it was shown for several countries and for a recent five-week period that the achieved credibility interval-based estimations of both denoised new infection counts and reproduction number provide practitioners with an efficient and robust tool for the actual and practical monitoring of the Covid19 pandemic. Such estimates are updated on a daily basis on the authors's web-pages. Automated and data-driven estimations of the hyperparameters λ_{R} and λ_{O} are under current investigations.

In an effort toward reproducible research and open science, Matlab codes implementing `PGdec`, `PGdual` for both Metropolis-Hastings and Gibbs versions are made publicly available at `https://github.com/gfort-lab`.

REFERENCES

[1] P. Abry, G. Fort, B. Pascal, and N. Pustelnik. Temporal evolution of the Covid19 pandemic reproduction number: Estimations from Proximal optimization to Monte Carlo sampling. Technical report, HAL, 2022.

[2] P. Abry et al. Spatial and temporal regularization to estimate COVID-19 reproduction number $R(t)$: Promoting piecewise smoothness via convex optimization. *PLOS One*, 15, 2020. e0237901.

[3] A. Ali and R.J. Tibshirani. The Generalized Lasso Problem and Uniqueness. *Electron. J. of Stat*, 13(2):2307 – 2347, 2019.

[4] J. Arino. Describing, modelling and forecasting the spatial and temporal spread of COVID-19–A short review. Technical report, arXiv:2102.02457, 2021.

[5] H. Artigas, B. Pascal, G. Fort, P. Abry, and N. Pustelnik. Credibility interval design for Covid19 reproduction number from nonsmooth Langevin-type Monte Carlo sampling. Technical report, hal-03371837, 2021.

[6] Y.F. Atchadé. A Moreau-Yosida approximation scheme for a class of high-dimensional posterior distributions. Technical report, arXiv: Statistics Theory, 2015.

[7] H. H. Bauschke and P.-L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.

[8] H. H Bauschke and P.-L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer International Publishing, 2017.

[9] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*, 7:434–455, 1998.

[10] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J Math Imaging Vis*, 40(1):120–145, 2011.

[11] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[12] K.S. Chan and C. J. Charles J. Geyer. Discussion: Markov Chains for Exploring Posterior Distributions. *Ann. Stat.*, 22(4):1747–1758, 1994.

[13] N. Chatterji, J. Diakonikolas, M. I. Jordan, and P. Bartlett. Langevin Monte Carlo without smoothness. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1716–1726, 2020.

[14] P.-L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke et al., editor, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer-Verlag, New York, 2011.

[15] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.*, 178:1505–1512, 2013.

[16] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.*, 28:365–382, 1990.

[17] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2018.

[18] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via Convex Optimization. *J. Mach. Learn. Res.*, 20:73:1–73:46, 2019.

[19] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann Appl Probab*, 27:1551 – 1587, 2017.

[20] A. Durmus, É. Moulines, and M. Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM J Imaging Sci*, 11:473–506, 2018.

[21] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797, 2018.

[22] M.A.T. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1150–1159, 2003.

[23] A. Flahault. COVID-19 cacophony: is there any orchestra conductor? *The Lancet*, 395(10229):1037, 2020.

[24] A. Gelman, W.R. Gilks, and G.O. Roberts. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann Appl Probab*, 7(1):110–120, 1997.

[25] A. Gelman and D.B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Stat Sci*, 7(4):457–472, 1992.

[26] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. and Mach. Intell.*, PAMI-6(6):721–741, 1984.

[27] G. Guzzetta et al. The impact of a nation-wide lockdown on COVID-19 transmissibility in Italy. arXiv:2004.12338 [q-bio.PE], 2020.

[28] J. Kent. Time-Reversible Diffusions. *Adv Appl Probab*, 10(4):819–835, 1978.

[29] Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani. Measurability of the epidemic reproduction number in data-driven contact networks. *Proc. Natl. Acad. Sci. U.S.A.*, 115:12680–12685, 2018.

[30] T.D. Luu, J. Fadili, and C. Chesneau. Sampling from Non-smooth Distributions Through Langevin Diffusion. *Methodol. Comput. Appl. Probab.*, 23:1173—-1201, 2021.

[31] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stat.*, 24(1):101–121, 1996.

[32] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.

[33] P. Moulin and J Liu. Analysis of multiresolution image denoising schemes using generalised Gaussian and complexity priors. *IEEE Trans. Inform. Theory*, 45, 1999.

[34] T. Obadia, R. Haneef, and P.-Y. Boëlle. The R0 package: A toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Inform Decis. Mak.*, 12:147, 2012.

[35] G. Parisi. Correlation functions and computer simulations. *Nucl. Phys. B*, 180(3):378–384, 1981.

[36] T. Park and G. Casella. The Bayesian Lasso. *J Am Stat Assoc*, 103(482):681–686, 2008.

[37] B. Pascal, P. Abry, N. Pustelnik, S. Roux, R. Gribonval, and P. Flandrin. Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low quality data. Technical report, arXiv 2109.09595, 2022.

[38] B. Pascal, N. Pustelnik, P. Abry, and J.-C. Pesquet. Block-coordinate proximal algorithms for scale-free texture segmentation. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018.

[39] N. Pustelnik, C. Chaux, and J.-C. Pesquet. Parallel proXimal algorithm for image restoration using hybrid regularization. *IEEE Trans. Image Process.*, 20:2450–2462, 2011.

[40] N. Pustelnik and L. Condat. Proximity operator of a sum of functions; application to depth map estimation. *IEEE Signal Process. Lett.*, 24(12):1827–1831, December 2017.

[41] F. Riccardo et al. Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic. medRxiv:2020.04.08.20056861, 2020.

[42] C. P. Robert. *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[43] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005.

[44] G. O. Roberts and J.S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann Appl Probab*, 16(4):2123 – 2139, 2006.

[45] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2:341 – 363, 1996.

[46] G.O. Roberts and O. Stramer. Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodol. Comput. Appl. Probab.*, 4:337–357, 2002.

[47] A. Salim and P. Richtarik. Primal Dual Interpretation of the Proximal Stochastic Gradient Langevin Algorithm. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3786–3796, 2020.

[48] A. Schreck, G. Fort, S. Le Corff, and É. Moulines. A Shrinkage-Thresholding Metropolis Adjusted Langevin Algorithm for Bayesian Variable Selection. *IEEE J. Selected Topics Signal Process.*, 10:366–375, 2016.

[49] R.N. Thompson et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356, 2019.

[50] P. van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math Biosci*, 180:29–48, 2002.

[51] J. Wallinga and P. Teunis. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *Am. J. Epidemiol.*, 160:509–516, 2004.

[52] Y.-L. Yu. On decomposing the proximal map. In *Proc. Ann. Conf. Neur. Inform. Proc. Syst.*, pages 91–99, Oaxaca, Mexico, 2013.

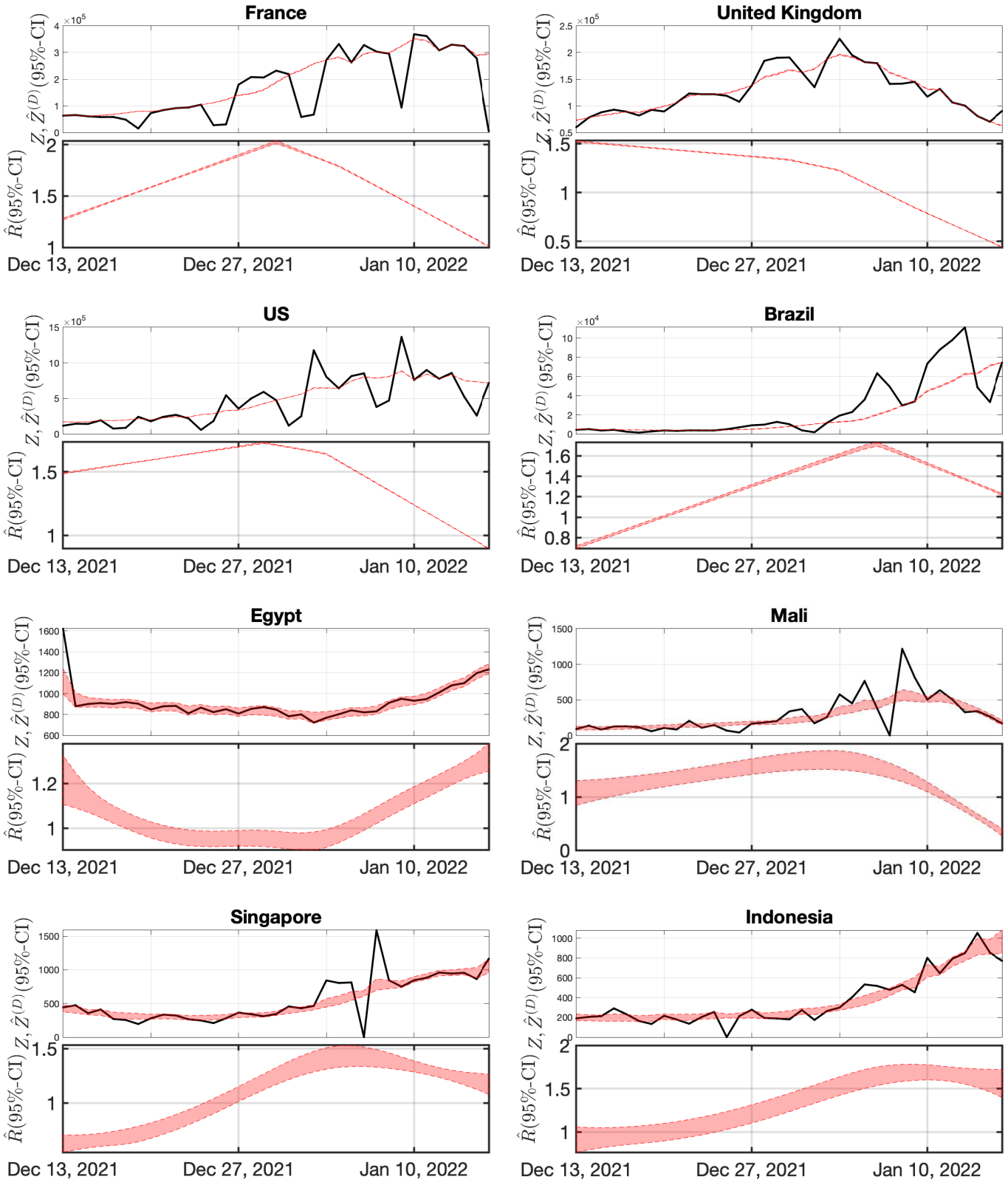

Fig. 2: `Gibbs PGdual O` **sampler-based estimation of the time evolution of** R **for a recent 5-week time period and for several countries.** Top rows: Raw daily new infections counts $\mathbf{Z}$ (black) and estimates of the denoised counts $\mathbf{Z}^{(D)}$, obtained by subtracting the 95%-Credibility interval estimates of the outliers to the raw new infection counts $\mathbf{Z}$; Bottom rows: 95%-Credibility interval estimated of R_t.

Supplementary material

VIII. Proof of proposition 1

A. Notations

Let $\mathcal{D}_{\mathbf{Z}}$ be the subset of $(\mathbb{R}_+)^T \times \mathbb{R}^T$ given by (2). To shorten the notations, we will write $p_t(\boldsymbol{\theta})$ instead of $p(\mathsf{R}_t, \mathsf{O}_t|\mathbf{Z}_{t-\tau_\phi:t-1})$ (see (3)):

$$p_t(\boldsymbol{\theta}) := \mathsf{R}_t \Phi_t^{\mathbf{Z}} + \mathsf{O}_t, \qquad \Phi_t^{\mathbf{Z}} := \sum_{u=1}^{\tau_\phi} \Phi_u \mathsf{Z}_{t-u}; \tag{30}$$

and π instead of $\pi(\cdot|\mathbf{Z})$ (see (6)) . Observe that $-\ln \pi(\boldsymbol{\theta})$ is equal to $+\infty$ for $\boldsymbol{\theta} \notin \mathcal{D}_{\mathbf{Z}}$ and for $\theta \in \mathcal{D}_{\mathbf{Z}}$

$$-\ln \pi(\boldsymbol{\theta}) = C_\pi + \sum_{t=1}^{T} \{p_t(\boldsymbol{\theta}) - \mathsf{Z}_t 1_{\mathsf{Z}_t>0} \ln p_t(\boldsymbol{\theta})\} + \lambda_{\mathsf{R}} \|\mathsf{D}_2 \mathbf{R}\|_1 + \lambda_{\mathsf{O}} \|\mathbf{O}\|_1;$$

for some normalizing constant C_π. By convention $0 \ln 0 = 0$. Define the $T \times T$ invertible matrix $\overline{\mathsf{D}}_2$ and compute its inverse:

$$\overline{\mathsf{D}}_2 := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -2 & 1 & 0 & \cdots & 0 \\ & & \mathsf{D}_2 & & \end{bmatrix};$$

$$\overline{\mathsf{D}}_2^{-1} := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 2 & 1 & 0 & \cdots & 0 \\ \cdots & & & & \\ T & (T-1) & \cdots & 2 & 1 \end{bmatrix}.$$

Finally, define the criterion

$$\mathcal{C}(\widetilde{\mathbf{R}}, \mathbf{O}) := \lambda_{\mathsf{R}} \|\widetilde{\mathbf{R}}_{3:T}\|_1 + \lambda_{\mathsf{O}} \|\mathbf{O}\|_1 + \sum_{t=1}^{T} \left\{ p_t(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) - \mathsf{Z}_t 1_{\mathsf{Z}_t>0} \ln p_t(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) \right\}, \tag{31}$$

for $(\widetilde{\mathbf{R}}, \mathbf{O})$ in the set $\widetilde{\mathcal{D}}_{\mathbf{Z}} := \{(\widetilde{\mathbf{R}}, \mathbf{O}) \text{ s.t. } (\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) \in \mathcal{D}_{\mathbf{Z}}\}$ and $+\infty$ otherwise. We have for any $(\widetilde{\mathbf{R}}, \mathbf{O}) \in \widetilde{\mathcal{D}}_{\mathbf{Z}}$

$$\mathcal{C}(\widetilde{\mathbf{R}}, \mathbf{O}) = -\ln \pi(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) - C_\pi. \tag{32}$$

B. Existence of a MAP

We start with two lower bounds on the criterion $\mathcal{C}$ which will help us to study its behavior on some boundaries of $\widetilde{\mathcal{D}}_{\mathbf{Z}}$. ► **Lower bounds on $\mathcal{C}$.** Let $(\widetilde{\mathbf{R}}, \mathbf{O}) \in \widetilde{\mathcal{D}}_{\mathbf{Z}}$. Then

$$\mathcal{C}(\widetilde{\mathbf{R}}, \mathbf{O}) \geq -\sum_{t=1}^{T} \ln(\mathsf{Z}_t!) + \lambda_{\mathsf{R}} \|(\widetilde{\mathbf{R}})_{3:T}\|_1 + \lambda_{\mathsf{O}} \|\mathbf{O}\|_1, \tag{33}$$

and for any $\tau \in \{1, \cdots, T\}$,

$$\mathcal{C}(\widetilde{\mathbf{R}}, \mathbf{O}) \geq \left\{ p_\tau(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) - \mathsf{Z}_\tau 1_{\mathsf{Z}_\tau>0} \ln p_\tau(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) \right\} - \sum_{t \neq \tau} \ln(\mathsf{Z}_t!). \tag{34}$$

Proof. For any $p > 0$ and $z \in \mathbb{N}$, $p^z \exp(-p)/z! \in (0, 1)$ thus implying that $z \ln p - p - \ln(z!) \leq 0$. When $p = z = 0$, then $z 1_{z>0} \ln p - p = 0$ and $\ln(z!) = 0$; hence we have

$$p_t(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) - \mathsf{Z}_t 1_{\mathsf{Z}_t>0} \ln p_t(\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}, \mathbf{O}) \geq -\ln(\mathsf{Z}_t!),$$

from which we obtain (33) and (34). □

► **Behavior of $\mathcal{C}$ on some boundaries of $\widetilde{\mathcal{D}}_{\mathbf{Z}}$.** Assume that there exist $t_\star < t_{\star\star}$ in $\{1, \ldots, T\}$ such that $\Phi_{t_\star}^{\mathbf{Z}} > 0$ and $\Phi_{t_{\star\star}}^{\mathbf{Z}} > 0$.

1) For any sequence $(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) \in \widetilde{\mathcal{D}}_{\mathbf{Z}}$ s.t. $\lim_n \|\widetilde{\mathbf{R}}^n\|_1 + \lim_n \|\mathbf{O}^n\|_1 = +\infty$, we have $\lim_n \mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$.
2) Let $t \in \{1, \ldots, T\}$ such that $\mathsf{Z}_t > 0$. For any sequence $(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) \in \widetilde{\mathcal{D}}_{\mathbf{Z}}$ s.t. $\lim_n (\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}^n)_t \Phi_t^{\mathbf{Z}} + \mathbf{O}_t^n = 0$, we have $\lim_n \mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$.

Proof. First statement. Let $\{(\widetilde{\mathbf{R}}^n, \mathbf{O}^n), n \geq 0\}$ be a sequence in $\widetilde{\mathcal{D}}_{\mathbf{Z}}$. For the discussions below, remember that this implies that

$$\lim_n \left((\overline{\mathsf{D}}_2^{-1} \widetilde{\mathbf{R}}^n)_t \Phi_t^{\mathbf{Z}} + \mathbf{O}_t^n \right) \geq 0. \tag{35}$$

We distinguish two cases. First, either $\|\mathbf{O}^n\|_1$ tends to infinity or $\|\mathbf{R}^n_{3:T}\|_1$ tends to infinity. In the second case, these two norms are assumed bounded but $\|\mathbf{R}^n_{1:2}\|_1$ tends to infinity.

- `First case.` Assume first $\lim_n\{\|\widetilde{\mathbf{R}}^n_{3:T}\|_1 + \|\mathbf{O}^n\|_1\} = +\infty$. Then, from (33), we have $\lim_n \mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$.
- `Second case.` Now consider the case when

$$\sup_n \left(\|\widetilde{\mathbf{R}}^n_{3:T}\|_1 \sup_n \|\mathbf{O}^n\|_1 \right) < \infty, \quad \lim_n \|\widetilde{\mathbf{R}}^n_{1:2}\|_1 = +\infty.$$

By definition of $\overline{\mathsf{D}}_2^{-1}$, we have for any $t \in \{1, \dots, T\}$,

$$(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n)_t = t\widetilde{\mathbf{R}}^n_1 + (t-1)\widetilde{\mathbf{R}}^n_2 + \sum_{k=3}^{t}(t-k+1)\widetilde{\mathbf{R}}^n_k; \tag{36}$$

by convention, the last term in the RHS is zero when $t = 1, 2$. Under the assumptions of this second case, $\sup_n |\sum_{k=3}^t (t-k+1)\widetilde{\mathbf{R}}^n_k| < \infty$ for any $t \in \{3, \dots, T\}$. We prove that either $(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n)_{t_\star}$ tends to infinity, or $(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n)_{t_{\star\star}}$ tends to infinity – which will imply by (34) that the criterion tends to infinity.

- `Subcase 2A.` Assume that $\lim_n\{t_\star\widetilde{\mathbf{R}}^n_1 + (t_\star - 1)\widetilde{\mathbf{R}}^n_2\} = +\infty$ (observe that this limit can not be $-\infty$, as a consequence of the assumptions of Case 2, (36) and (35)).

Apply (36) with $t = t_\star$; this yields $\lim_n(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n)_{t_\star} = +\infty$. Since $\Phi^{\mathbf{Z}}_{t_\star > 0}$ then $\lim_n p_{t_\star}(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$ (see (30)) which implies that

$$\lim_n \left\{ p_{t_\star}(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n, \mathbf{O}^n) - \mathsf{Z}_{t_\star} \ln p_{t_\star}(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n, \mathbf{O}^n) \right\} = +\infty$$

and then $\lim_n \mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$ by (34).

- `Subcase 2B.` Assume that $\sup_n |t_\star\widetilde{\mathbf{R}}^n_1 + (t_\star - 1)\widetilde{\mathbf{R}}^n_2| < \infty$. Then, necessarily $\lim_n |\widetilde{\mathbf{R}}^n_2| = +\infty$ (otherwise, it is the Subcase 2A). We write

$$\begin{aligned}(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n)_{t_{\star\star}} - \sum_{k=3}^{t_{\star\star}}(t_{\star\star}-k+1)\widetilde{\mathbf{R}}^n_k & \\ = t_{\star\star}\widetilde{\mathbf{R}}^n_1 + (t_{\star\star}-1)\widetilde{\mathbf{R}}^n_2 & \\ = \frac{t_{\star\star}}{t_\star}\left(t_\star\widetilde{\mathbf{R}}^n_1 + (t_\star - 1)\widetilde{\mathbf{R}}^n_2\right) + \frac{t_{\star\star}}{t_\star}(1 - \frac{t_\star}{t_{\star\star}})\widetilde{\mathbf{R}}^n_2. &\end{aligned}$$

Since $t_\star < t_{\star\star}$, this equality and (35) imply that $\lim_n \widetilde{\mathbf{R}}^n_2 = +\infty$. Therefore, since $t_\star < t_{\star\star}$, we have $\lim_n(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n)_{t_{\star\star}} = +\infty$. We then conclude, along the same lines as in Subcase 2A, that $\lim_n \mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$.

Second statement. Let $\{(\widetilde{\mathbf{R}}^n, \mathbf{O}^n), n \geq 0\}$ be a sequence in $\widetilde{\mathcal{D}}_{\mathbf{Z}}$ and τ such that $\mathsf{Z}_\tau > 0$. By (34), we have

$$\mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) + \sum_{t \neq \tau} \ln(\mathsf{Z}_t!) \geq p_\tau(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n, \mathbf{O}^n) - \mathsf{Z}_\tau \ln p_\tau(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n, \mathbf{O}^n).$$

The RHS tends to $+\infty$ since $\lim_n p_\tau(\overline{\mathsf{D}}_2^{-1}\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = 0$ and $\mathsf{Z}_\tau > 0$ by assumptions, hence, $\lim_n \mathcal{C}(\widetilde{\mathbf{R}}^n, \mathbf{O}^n) = +\infty$. □

► **Conclusion: a MAP exists.** Assume that there exist $t_\star < t_{\star\star}$ in $\{1, \dots, T\}$ such that $\Phi^{\mathbf{Z}}_{t_\star} > 0$ and $\Phi^{\mathbf{Z}}_{t_{\star\star}} > 0$. Then $-\ln \pi$ possesses at least one minimizer in $\mathcal{D}_{\mathbf{Z}}$.

Proof. Consider the point $(\mathbf{R}^0, \mathbf{O}^0)$ given by $\mathbf{R}^0 := (1, \dots, 1)^\top$ and $\mathbf{O}^0 := (1, \dots, 1)^\top$. Then, $(\mathbf{R}^0, \mathbf{O}^0) \in \mathcal{D}_{\mathbf{Z}}$. Set $M^0 := -\ln \pi(\mathbf{R}^0, \mathbf{O}^0)$.

The goal of the proof below is to build a closed bounded set $\mathcal{K}$ in $\mathcal{D}_{\mathbf{Z}}$ such that outside $\mathcal{K}$, $-\ln \pi \geq 1 + M^0$. This implies that $(\mathbf{R}^0, \mathbf{O}^0) \in \mathcal{K}$. Since $-\ln \pi$ is continuous on $\mathcal{D}_{\mathbf{Z}}$, it reaches its minimum on the compact subset $\mathcal{K}$, and this minimum is upper bounded by M^0. Hence, this minimizer is also a global minimizer. Let us define $\mathcal{K}$. We have

$$\lambda_{\min}\|\widetilde{\mathbf{R}}\|^2 \leq \|\mathsf{D}_2^{-1}\widetilde{\mathbf{R}}\|^2 = \|\mathbf{R}\|^2 \leq \lambda_{\max}\|\widetilde{\mathbf{R}}\|^2$$

where $\lambda_{\min}$ (resp. $\lambda_{\max}$) is the minimal (resp. maximal) eigenvalue of $\overline{\mathsf{D}}_2^{-\top}\overline{\mathsf{D}}_2^{-1}$; they are positive and finite. Consequently, by setting $\widetilde{\mathbf{R}}^n := \overline{\mathsf{D}}_2\mathbf{R}^n$, we have $\|(\widetilde{\mathbf{R}}^n, \mathbf{O}^n)\|_\ell \to +\infty$ iff $\|(\mathbf{R}^n, \mathbf{O}^n)\|_\ell \to +\infty$ for $\ell = 1, 2$ since the norms are equivalent on $\mathbb{R}^{2T}$. This property, the coercivity property (statement 1), and the equality (32) imply that $\lim_n -\ln \pi(\mathbf{R}^n, \mathbf{O}^n) = +\infty$ for any $\mathcal{D}_{\mathbf{Z}}$-valued sequence $\{(\mathbf{R}^n, \mathbf{O}^n), n \geq 0\}$ such that $\lim_n \|\mathbf{R}^n\| + \lim_n \|\mathbf{O}\|^n = +\infty$. As a consequence, there exists C_{1+M^0} such that

$$(\mathbf{R}, \mathbf{O}) \in \mathcal{D}_{\mathbf{Z}}, \|\mathbf{R}\| + \|\mathbf{O}\| > C_{1+M_0} \Longrightarrow -\ln \pi(\mathbf{R}, \mathbf{O}) \geq 1 + M^0.$$

Similarly, there exists $c_{1+M_0} > 0$ such that

$$(\mathbf{R}, \mathbf{O}) \in \mathcal{D}_{\mathbf{Z}}, \mathbf{R}_t\Phi^{\mathbf{Z}}_t + \mathbf{O}_t < c_{1+M^0} \text{ for some } t \text{ s.t. } \mathsf{Z}_t > 0 \Longrightarrow -\ln \pi(\mathbf{R}, \mathbf{O}) \geq 1 + M^0.$$

Consequently, we define

$$\mathcal{K} := \mathcal{D}_{\mathbf{Z}} \cap \{\boldsymbol{\theta} : \|\mathbf{R}\| + \|\mathbf{O}\| \leq C_{1+M_0}\} \cap \{\boldsymbol{\theta} : \mathbf{R}_t \Phi_t^{\mathbf{Z}} + \mathbf{O}_t \geq c_{1+M^0} \text{ for } t \text{ s.t. } \mathsf{Z}_t > 0\}.$$

□

C. About the uniqueness of the MAP

We just proved that there exists a compact subset of the interior of $\mathcal{D}_{\mathbf{Z}}$ that contains a minimizer of $-\ln \pi$. Let $\boldsymbol{\theta}^\star = (\mathbf{R}^\star, \mathbf{O}^\star)$ be a minimizer.

• **One or uncountably many.** The function $-\ln \pi$ is convex and finite on a convex set: hence, given a second minimizer $\boldsymbol{\theta}^{\star\star}$, $\mu\boldsymbol{\theta}^\star + (1-\mu)\boldsymbol{\theta}^{\star\star}$ is also a minimizer, whatever $\mu \in [0,1]$.

• **Same intensity, data fidelity term, and penalty term.** Let $f_{\mathbf{Z}}$ and g be defined by (7). Following the same lines as in [37] where the strict convexity of the Kullback-Leibler term $f_{\mathbf{Z}}$ is the key ingredient, it can be proved that $p_t(\boldsymbol{\theta}^\star) = p_t(\boldsymbol{\theta}^{\star\star})$ for any $t \in \{1, \dots, T\}$ and thus $f_{\mathbf{Z}}(\boldsymbol{\theta}^\star) = f_{\mathbf{Z}}(\boldsymbol{\theta}^{\star\star})$; since $-\ln \pi(\boldsymbol{\theta}^\star) = -\ln \pi(\boldsymbol{\theta}^{\star\star})$ since both points minimize $-\ln \pi$, we have $g(\boldsymbol{\theta}^\star) = g(\boldsymbol{\theta}^{\star\star})$.

• **Sign conditions.** Set $\mathrm{sign}(a) = 1$ when $a > 0$, $\mathrm{sign}(a) = -1$ when $a < 0$ and $\mathrm{sign}(a) = 0$ when $a = 0$. For A, B in $\mathbb{R}$, and $\mu > 0$, we have $|A + \mu B| = |A| + \mu \mathrm{sign}(A)\,\mathrm{sign}(B)|B| 1_{A\neq 0} + \mu |B| 1_{A=0} + \mu o(1)$ where $o(1)$ is a function satisfying $\lim_{\mu\to 0} o(1) = 0$. Hence, for $\boldsymbol{\tau} = \tau_{1:d}, \boldsymbol{\tau}' = \tau'_{1:d} \in \mathbb{R}^d$ and $\mu \in (0,1)$,

$$\|\boldsymbol{\tau} + \mu(\boldsymbol{\tau}' - \boldsymbol{\tau})\|_1 = (1-\mu)\,\|\boldsymbol{\tau}\|_1 + \mu \|\boldsymbol{\tau}'\|_1 + \mu \sum_{t=1}^{d} (\mathrm{sign}(\tau_t)\mathrm{sign}(\tau'_t) - 1)\,|\tau'_t| 1_{\tau_t \neq 0} + \mu\, o(1).$$

Set $\boldsymbol{\theta}^\mu := \boldsymbol{\theta}^\star + \mu(\boldsymbol{\theta}^{\star\star} - \boldsymbol{\theta}^\star)$; as proved above, $f_{\mathbf{Z}}(\boldsymbol{\theta}^\mu) = f_{\mathbf{Z}}(\boldsymbol{\theta}^\star)$. We prove that if the sign conditions do not hold, for μ small enough $g(\boldsymbol{\theta}^\mu) < g(\boldsymbol{\theta}^\star)$ which yields a contradiction since $\boldsymbol{\theta}^\star$ is a minimizer. By (**??**), we obtain

$$\|\mathbf{O}^\star + \mu(\mathbf{O}^{\star\star} - \mathbf{O}^\star)\|_1 - (1-\mu)\|\mathbf{O}^\star\|_1 - \mu\|\mathbf{O}^{\star\star}\|_1 = \mu \sum_{t=1}^{T} (\mathrm{sign}(\mathsf{O}_t^\star)\mathrm{sign}(\mathsf{O}_t^{\star\star}) - 1)\,|\mathsf{O}_t^{\star\star}| 1_{\mathsf{O}_t^\star \neq 0} + \mu\, o(1).$$

We have a similar expansion for $\|\mathsf{D}_2\mathbf{R}^\star + \mu(\mathsf{D}_2\mathbf{R}^{\star\star} - \mathsf{D}_2\mathbf{R}^\star)\|_1$. Since $g(\boldsymbol{\theta}^\star) = g(\boldsymbol{\theta}^{\star\star})$, this yields

$$\begin{aligned} g(\boldsymbol{\theta}^\mu) - g(\boldsymbol{\theta}^\star) = \mu \sum_{t=1}^{T-2} |(\mathsf{D}_2\mathbf{R}^{\star\star})_t|\,(\mathrm{sign}((\mathsf{D}_2\mathbf{R}^\star)_t)\ \mathrm{sign}((\mathsf{D}_2\mathbf{R}^{\star\star})_t) - 1)\, 1_{(\mathsf{D}_2\mathbf{R}^\star)t \neq 0} \\ + \mu \sum_{t=1}^{T} (\mathrm{sign}(\mathsf{O}_t^\star)\ \mathrm{sign}(\mathsf{O}_t^{\star\star}) - 1)\,|\mathsf{O}_t^{\star\star}| 1_{\mathsf{O}_t^\star \neq 0} + \mu\, o(1). \end{aligned}$$

For μ small enough, the RHS is negative when for some t, $\mathrm{sign}(\mathsf{O}_t^\star)\mathrm{sign}(\mathsf{O}_t^{\star\star}) = -1$ or when for some s, $\mathrm{sign}((\mathsf{D}_2\mathbf{R}^\star)_s)\mathrm{sign}((\mathsf{D}_2\mathbf{R}^{\star\star})_s) = -1$. If such, $g(\boldsymbol{\theta}^\mu) < g(\boldsymbol{\theta}^\star)$.

• **Sufficient conditions for uniqueness.** The proof is adapted from [3, Section 4]. Let $\boldsymbol{\theta}^{\star\star} = \boldsymbol{\theta}^\star + \boldsymbol{\omega}$ be another minimizer. Define

$$\mathsf{U} := \begin{bmatrix} \lambda_{\mathrm{R}} \mathsf{D}_2 & 0_{(T-2)\times T} \\ 0_{T\times T} & \lambda_{\mathrm{O}}\, \mathrm{I}_T \end{bmatrix} \in \mathbb{R}^{(2T-2)\times(2T)}.$$

The Fermat rule ([8, Theorem 16.2]) which characterizes optimality implies that zero is in the subdifferential of $-\ln \pi$ at $\boldsymbol{\theta}^\star$: there exists $\gamma(\boldsymbol{\theta}^\star) \in \mathbb{R}^{2T-2}$ such that $\nabla f_{\mathbf{Z}}(\boldsymbol{\theta}^\star) + \mathsf{U}^\top \gamma(\boldsymbol{\theta}^\star) = 0_{2T\times 1}$ where $\gamma(\boldsymbol{\theta}^\star)$ is the subgradient of the L^1-norm in $\mathbb{R}^{2T-2}$ evaluated at $\mathsf{U}\boldsymbol{\theta}^\star \in \mathbb{R}^{2T-2}$. Since $\nabla f_{\mathbf{Z}}(\boldsymbol{\theta}^\star)$ depends on $\boldsymbol{\theta}^\star$ through the p_t's which are constant on the set of the minimizers $\mathcal{M}$ (see above), and $\mathsf{U}^\top$ has full column rank, $\gamma(\boldsymbol{\theta}^\star)$ is the same whatever the minimizer $\boldsymbol{\theta}^\star$; it is denoted by $\gamma^\star$. Set $\mathcal{I} := \{j \in \{1, \dots, 2T-2\} : |\gamma_j^\star| < 1\}$. Observe that any minimizer is in the kernel K_1 of the matrix $\mathsf{U}_{\mathcal{I}}$ which, by definition, collects the rows of U indexed by $\mathcal{I}$; hence $\boldsymbol{\omega} \in \mathsf{K}_1$. In addition, $\boldsymbol{\omega}$ is in the kernel K_2 of the $T \times (2T)$ matrix $[\mathsf{diag}(\Phi_1^{\mathbf{Z}}, \dots \Phi_T^{\mathbf{Z}})\ \mathrm{I}_T]$, since all the p_t's are constant on $\mathcal{M}$. Therefore, if $\mathsf{K}_1 \cap \mathsf{K}_2 = \{0\}$, the MAP is unique.

IX. Proof of section III

A. Detailed proof of Theorem 10

Throughout the proof, we write A, U and $\bar{\mathsf{A}}$ as a shorthand notation for $\mathsf{A}_{i,j}$, $\mathsf{U}_{i,j}$ and $\bar{\mathsf{A}}_{i,j}$. Under the stated assumptions,

$$\bar{\mathsf{A}}^{-1} = \begin{bmatrix} \mathsf{U}^\top(\mathsf{U}\mathsf{U}^\top)^{-1} & \mathsf{A}^\top(\mathsf{A}\mathsf{A}^\top)^{-1} \end{bmatrix}. \tag{37}$$

We first focus on the gradient step in (18) leading to:

$$\bar{\mathsf{A}}\boldsymbol{\theta}_j - \gamma_j \bar{\mathsf{A}}^{-\top} \nabla_j f(\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{U}\boldsymbol{\theta}_j - \gamma_j (\mathsf{U}\mathsf{U}^\top)^{-1}\mathsf{U}\nabla_j f(\boldsymbol{\theta}) \\ \mathsf{A}\boldsymbol{\theta}_j - \gamma_j (\mathsf{A}\mathsf{A}^\top)^{-1}\mathsf{A}\nabla_j f(\boldsymbol{\theta}) \end{bmatrix}.$$

Second, for any $\boldsymbol{\tau} = \boldsymbol{\tau}_{1:d_j} \in \mathbb{R}^{d_j}$,

$$\operatorname{prox}_{\gamma_j \bar{g}_{i,j}}(\boldsymbol{\tau}) = \operatorname{prox}_{\gamma_j g_{i,j}(\mathsf{A}\bar{\mathsf{A}}^{-1}\cdot)}(\boldsymbol{\tau}) \tag{38}$$

$$= \begin{bmatrix} \boldsymbol{\tau}_{1:d_j - c_{i,j}} \\ \operatorname{prox}_{\gamma_j g_{i,j}}(\boldsymbol{\tau}_{d_j - c_{i,j}+1:d_j}) \end{bmatrix}, \tag{39}$$

since under the stated assumptions, we have

$$\mathsf{A}\bar{\mathsf{A}}^{-1} = \begin{bmatrix} 0_{c_{i,j} \times (d_j - c_{i,j})} & \mathrm{I}_{c_{i,j}} \end{bmatrix}.$$

Therefore,

$$\operatorname{prox}_{\gamma_j\, \bar{g}_{i,j}}\left(\bar{\mathsf{A}}\boldsymbol{\theta}_j - \gamma_j \bar{\mathsf{A}}^{-\top}\nabla_j f(\boldsymbol{\theta})\right) = \begin{bmatrix} \mathsf{U}\boldsymbol{\theta}_j - \gamma_j(\mathsf{U}\mathsf{U}^\top)^{-1}\mathsf{U}\nabla_j f(\boldsymbol{\theta}) \\ \operatorname{prox}_{\gamma_j\, g_{i,j}}\left(\mathsf{A}\boldsymbol{\theta}_j - \gamma_j(\mathsf{A}\mathsf{A}^\top)^{-1}\mathsf{A}\nabla_j f(\boldsymbol{\theta})\right) \end{bmatrix}.$$

Now, let us apply $\bar{\mathsf{A}}^{-1}$; by (37), we have

$$\begin{aligned}\bar{\mathsf{A}}^{-1} \operatorname{prox}_{\gamma_j\, g_{i,j}(\mathsf{A}\bar{\mathsf{A}}^{-1}\cdot)}\left(\bar{\mathsf{A}}\boldsymbol{\theta}_j - \gamma_j \bar{\mathsf{A}}^{-\top}\nabla_j f(\boldsymbol{\theta})\right) &= \mathsf{U}^\top(\mathsf{U}\mathsf{U}^\top)^{-1}\left(\mathsf{U}\boldsymbol{\theta}_j - \gamma_j(\mathsf{U}\mathsf{U}^\top)^{-1}\mathsf{U}\nabla_j f(\boldsymbol{\theta})\right) \\ &\quad + \mathsf{A}^\top(\mathsf{A}\mathsf{A}^\top)^{-1}\operatorname{prox}_{\gamma_j\, g_{i,j}}\left(\mathsf{A}\boldsymbol{\theta}_j - \gamma_j(\mathsf{A}\mathsf{A}^\top)^{-1}\mathsf{A}\nabla_j f(\boldsymbol{\theta})\right).\end{aligned}$$

Finally, since $\bar{\mathsf{A}}^{-1}\bar{\mathsf{A}} = \mathrm{I}_{d_j}$, we have

$$\mathsf{U}^\top(\mathsf{U}\mathsf{U}^\top)^{-1}\mathsf{U} + \mathsf{A}^\top(\mathsf{A}\mathsf{A}^\top)^{-1}\mathsf{A} = \mathrm{I}_{d_j}$$

and this concludes the proof of (24).
When $\mathsf{A}\mathsf{A}^\top = \mathrm{I}_{c_{i,j}}$, we have $\widetilde{\Omega}_{i,j} = \mathsf{A}^\top\mathsf{A}$ and $\mathsf{A}\widetilde{\Omega}_{i,j} = \mathsf{A}$. When $\mathsf{U}\mathsf{U}^\top = \mathrm{I}_{d_j - c_{i,j}}$, we have $\Omega_{i,j} = \mathsf{U}\mathsf{U}^\top$ and $\mathrm{I}_{d_j} - \Pi_{i,j} = \mathsf{U}^\top\mathsf{U}$; this yields

$$(\mathrm{I}_{d_j} - \Pi_{i,j})\Omega_{i,j} = \mathsf{U}^\top\mathsf{U}\mathsf{U}^\top\mathsf{U} = \mathsf{U}^\top\mathsf{U} = \mathrm{I}_{d_j} - \Pi_{i,j},$$

leading to the final result of Theorem 10.

B. Detailed proof of Proposition 13

Let a finite set $\mathcal{S}$ of indices. For any $\boldsymbol{\theta} \in \mathcal{D}$, let $\{\rho_\iota(\boldsymbol{\theta}), \iota \in \mathcal{S}\}$ be a weight function: $\sum_{\iota\in\mathcal{S}} \rho_\iota(\boldsymbol{\theta}) = 1$ and $\rho_\iota(\boldsymbol{\theta}) \geq 0$. Finally, for any $\iota \in \mathcal{S}$, let $q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}')\mathrm{d}\theta'$ be a Markov transition with respect to the Lebesgue measure on $\mathbb{R}^d$. `PGdec` and `PGdual` are special instances of algorithm 4. They correspond to the case $\mathcal{S} := \{(i_1, \ldots, i_J), i_j \in \{1, \ldots, I_j\}\}$; $\rho_\iota(\boldsymbol{\theta}) = 1/(I_1 I_2 \ldots I_J)$

Algorithm 4: General Blockwise Metropolis-Hastings.

Data: $N_{\max} \in \mathbb{N}_\star$, $\boldsymbol{\theta}^0 \in \mathcal{D}$
Result: A $\mathcal{D}$-valued sequence $\{\boldsymbol{\theta}^n, n \in [N_{\max}]\}$

for $n = 0, \ldots, N_{\max} - 1$ **do**
- Sample $\iota \in \mathcal{S}$ with distribution $\{\rho_i(\boldsymbol{\theta}^n), i \in \mathcal{S}\}$;
- Draw $\boldsymbol{\theta}^{n+1/2} \sim q_\iota(\boldsymbol{\theta}^n, \cdot)$;
- Set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+1/2}$ with probability $\alpha_\iota(\boldsymbol{\theta}^n, \boldsymbol{\theta}^{n+1/2})$

$$\alpha_\iota(x, y) := 1 \wedge \frac{\pi(y)}{\pi(x)}\frac{\rho_\iota(y)}{\rho_\iota(x)}\frac{q_\iota(y, x)}{q_\iota(x, y)}$$

  and $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$ otherwise.

for any $\boldsymbol{\theta}$; and to

$$q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') = \prod_{j=1}^{J} q_{i_j,j}(\boldsymbol{\theta}, \boldsymbol{\theta}'_j), \quad \boldsymbol{\theta} \in \mathcal{D}, \boldsymbol{\theta}' \in \mathcal{D}$$

where $\iota = (i_1, \ldots, i_J)$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_J)$.
Claim1. Assume: **[B1]** for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D}$, there exists $\iota \in \mathcal{S}$ such that $\rho_\iota(\boldsymbol{\theta})q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \wedge \rho_\iota(\boldsymbol{\theta}')q_\iota(\boldsymbol{\theta}', \boldsymbol{\theta}) > 0$; **[B2]** π is continuous on $\mathcal{D}$; **[B3]** for any compact set K of $\mathcal{D}$, $\inf_{K\times K}\sum_{\iota\in\mathcal{S}} \rho_\iota q_\iota > 0$. Then the sequence $\{\boldsymbol{\theta}^n, n \geq 0\}$ obtained by algorithm 4 is a Markov chain, taking values in $\mathcal{D}$. It is ϕ-irreducible, strongly aperiodic and π is its unique invariant distribution.

Proof. • $\boldsymbol{\theta}^n \in \mathcal{D}$ *for any* n. The proof is by induction on n. This property holds true for $n = 0$. Assume that $\boldsymbol{\theta}^n \in \mathcal{D}$. If $\boldsymbol{\theta}^{n+1/2} \notin \mathcal{D}$, then $\pi(\boldsymbol{\theta}^{n+1/2}) = 0$ and $\alpha_i(\boldsymbol{\theta}^n, \boldsymbol{\theta}^{n+1/2}) = 0$, so that $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$ and $\boldsymbol{\theta}^{n+1}$ is in $\mathcal{D}$. This concludes the induction.
• π *is an invariant probability measure.* Conditionally to ι and $\boldsymbol{\theta}^n$, the distribution of $\boldsymbol{\theta}^{n+1}$ is

$$P_\iota(\boldsymbol{\theta}^n, \mathrm{d}\boldsymbol{\theta}') := \delta_{\boldsymbol{\theta}^n}(\mathrm{d}\boldsymbol{\theta}')\left(1 - \int_{\mathbb{R}^d} \alpha_\iota(\boldsymbol{\theta}^n, \tau)q_\iota(\boldsymbol{\theta}^n, \tau)\mathrm{d}\tau\right) + \alpha_\iota(\boldsymbol{\theta}^n, \boldsymbol{\theta}')q_\iota(\boldsymbol{\theta}^n, \boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}' ;$$

$\delta_x(\mathrm{d}\boldsymbol{\theta}')$ denotes the Dirac mass at x. Conditionally to $\boldsymbol{\theta}^n$, the distribution of ι is $\{\rho_i(\boldsymbol{\theta}^n), i \in \mathcal{S}\}$. Hence the conditional distribution of $\boldsymbol{\theta}^{n+1}$ given $\boldsymbol{\theta}^n$ is

$$P_\star(\boldsymbol{\theta}^n, \mathrm{d}\boldsymbol{\theta}') := \sum_{i\in\mathcal{S}} \rho_i(\boldsymbol{\theta}^n) P_i(\boldsymbol{\theta}^n, \mathrm{d}\boldsymbol{\theta}') \; .$$

Following the sames lines as in [43, Theorem 7.2] (details are omitted), the *detailed balance condition* with π can be established

$$\pi(\boldsymbol{\theta}) \sum_{i\in\mathcal{S}} \rho_i(\boldsymbol{\theta}) \alpha_i(\boldsymbol{\theta}, \boldsymbol{\theta}') q_i(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}') \sum_{i\in\mathcal{S}} \rho_i(\boldsymbol{\theta}') \alpha_i(\boldsymbol{\theta}', \boldsymbol{\theta}) q_i(\boldsymbol{\theta}', \boldsymbol{\theta}) \; ;$$

hence π is invariant for $P_\star$.

• *Irreducibility.* By [B1], the chain is ϕ-irreducible (see [31, Lemma 1.1.]).

• *Aperiodicity.* Let us prove that the compact sets are 1-small and the chain is aperiodic; the proof is on the same lines as the proof of [31, Lemma 1.2.]. Let K be a compact set in $\mathcal{D}$. Since $\pi(x) < \infty$ on $\mathcal{D}$ then $\sup_K \pi < \infty$ by [B2]. For any measurable set $A \subseteq K$ and any $\boldsymbol{\theta} \in K$, it holds

$$\begin{aligned} P_\star(\boldsymbol{\theta}, A) &\geq \sum_{\iota\in\mathcal{S}} \rho_\iota(\boldsymbol{\theta}) \int_A q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \alpha_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}' \\ &\geq \int_A \sum_{\iota\in\mathcal{S}} \frac{\rho_\iota(\boldsymbol{\theta}) q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}')} \wedge \frac{\rho_\iota(\boldsymbol{\theta}') q_\iota(\boldsymbol{\theta}', \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}' \\ &\geq \frac{\inf_{K\times K} \sum_{\iota\in\mathcal{S}} \rho_\iota q_\iota}{\sup_K \pi} \int_A \pi(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'. \end{aligned}$$

The RHS is positive by [B3] and this proves that K is 1-small and the chain is aperiodic.

• *Unique invariant probability distribution.* Finally, [32, Propositions 10.1.1. and 10.4.4] prove that π is the unique invariant distribution. □

Claim 2. Both `PGdec` and `PGdual` satisfy [B1,B2,B3]. The `PGdec` and `PGdual` chains are positive Harris-recurrent Markov chains: they satisfy a strong law of large numbers for any initial value in $\mathcal{D}$.

Proof. • *Both algorithms satisfy [B1].* For both algorithms, $\rho_\iota(\boldsymbol{\theta}) = 1/(I_1 \cdots I_J)$ and $q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is proportional to

$$\prod_{j=1}^{J} \exp\left(-0.5(\boldsymbol{\theta}'_j - \mu_{i_j,j}(\boldsymbol{\theta}))^\top \mathsf{C}_{i_j,j}^{-1} (\boldsymbol{\theta}'_j - \mu_{i_j,j}(\boldsymbol{\theta}))\right)$$

where $\iota = (i_1, \cdots, i_J)$ and μ is $\boldsymbol{\mu}^{\mathrm{PGdec}}$ or $\boldsymbol{\mu}^{\mathrm{PGdual}}$. Therefore, since $\mu_{i,j}(\tau) < \infty$ for any $\tau \in \mathcal{D}$, we have $q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \wedge q_\iota(\boldsymbol{\theta}', \boldsymbol{\theta}) > 0$ for any $\iota \in \mathcal{S}$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D}$.

• *Both algorithms satisfy [B3].* For any compact set K of $\mathcal{D}$, we have $\sup_K \|\mu_{i,j}\| < \infty$; in addition, μ is a continuous function on $\mathcal{D}$ (the function f is continuously differentiable and the proximal operator is continuous by [8, Proposition 12.28]. Hence $\inf_{K\times K} q_\iota > 0$ and [B3] holds.

• *Positive Harris recurrence.* From Claim 1, the `PGdec` Markov chain and the `PGdual` one are positive recurrent (they are ϕ-irreducible with an invariant distribution, and recurrent by [32, Proposition 10.4.4]). Following the same lines as in [44, Theorem 8], we prove that the chain is Harris recurrent by showing that for any measurable set A such that $\int_A \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1$ and any $\boldsymbol{\theta} \in \mathcal{D}$, $\mathbb{P}_{\boldsymbol{\theta}}(\tau_A < \infty) = 1$ where τ_A is the return-time to the set A ([44, Theorem 6(v)]); here $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the probability on the canonical space of the Markov chain with initial distribution the Dirac mass at $\boldsymbol{\theta}$ and with kernel $P_\star$. Let A be a measurable subset of $\mathcal{D}$ such that $\int_A \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1$. Let $\boldsymbol{\theta} \in \mathcal{D}$. We write the kernel $P_\star$ as follows

$$P_\star(\boldsymbol{\theta}, A) = (1 - r(\boldsymbol{\theta})) M(\boldsymbol{\theta}, A) + r(\boldsymbol{\theta}) \delta_{\boldsymbol{\theta}}(A),$$

where $r(\boldsymbol{\theta}) := 1 - \sum_{\iota\in\mathcal{S}} \rho_\iota(\boldsymbol{\theta}) \int_{\mathcal{D}} q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \alpha_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'$, and

$$M(\boldsymbol{\theta}, A) := (1 - r(\boldsymbol{\theta}))^{-1} \sum_{\iota\in\mathcal{S}} \rho_\iota(\boldsymbol{\theta}) \int_A q_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \alpha_\iota(\boldsymbol{\theta}, \boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'.$$

Hence, $P_\star(\boldsymbol{\theta}, \cdot)$ is a mixture of two distributions: a Dirac mass at $\boldsymbol{\theta}$ and $M(\boldsymbol{\theta}, \cdot)$. Since $\int_{A^c} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0$ (here, $A^c := \mathcal{D} \setminus A$), then the Lebesgue measure of A^c is 0. This implies that $M(\boldsymbol{\theta}, A^c) = 0$ and $M(\boldsymbol{\theta}, A) = 1$. It holds

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}}(\tau_A = +\infty) &= \mathbb{E}_{\boldsymbol{\theta}}[1_{X_1 \notin A} \mathbb{P}_{X_1}(\tau_A = +\infty)] \\ &= \mathbb{E}_{\boldsymbol{\theta}}[1_{X_1 \in A^c} \mathbb{P}_{X_1}(\tau_A = +\infty)] \\ &= r(\boldsymbol{\theta}) \, \mathbb{P}_{\boldsymbol{\theta}}(\tau_A = +\infty); \end{aligned}$$

indeed, starting from $\boldsymbol{\theta}$, the chain can not reach A^c when the kernel $M(\boldsymbol{\theta}, \cdot)$ is selected; and remains at $\boldsymbol{\theta}$ when this kernel is not selected. Since $r(\boldsymbol{\theta}) < 1$ (otherwise the chain can not be ϕ-irreducible), we have $\mathbb{P}_{\boldsymbol{\theta}}(\tau_A = +\infty) = 0$. This concludes the

proof.

• `Strong Law of Large numbers`. A positive Harris recurrent chain satisfies a strong law of large numbers whatever the initial value in $\mathcal{D}$ ([32, Theorem 17.0.1]). □