

Méthodes de Monte Carlo proximales pour le suivi de pandémies

Mots-clés– Echantillonnage Monte Carlo, Chaînes de Markov, Optimisation convexe par méthodes proximales, Applications en épidémiologie.

Localisation: LAAS-CNRS Laboratoire d'Analyse et Architecture des Systèmes, Toulouse, France.

Encadrants : Gersende Fort (CNRS, LAAS)

email : gersende.fort@cnrs.fr

web : <http://perso.math.univ-toulouse.fr/gfort/>

Juliette Chevallier (INSA Toulouse, IMT)

email : juliette.chevallier@math.univ-toulouse.fr

web : <https://juliette-chevallier.pages.math.cnrs.fr/>

Dates: 4 à 6 mois entre mi février et fin août 2026.

Contexte. Les méthodes de Monte Carlo sont des algorithmes qui produisent une suite de points $\{X_k, k \geq 0\}$ approchant une loi de probabilité π [5]. Lorsque celle-ci est trop complexe pour que des statistiques d'intérêt, telles que la moyenne ou la variance ou les quantiles de ses marginales, puissent être obtenues par des formules explicites, les méthodes de Monte Carlo fournissent une approximation numérique

$$\int h(x)\pi(dx) \approx \frac{1}{K} \sum_{k=1}^K h(X_k). \quad (1)$$

Ces points peuvent être à dynamique markovienne : la construction du point X_{k+1} dépend des points précédents $(X_{k-r}, \dots, X_{k-1}, X_k)$ et d'un phénomène aléatoire indépendant; on parle alors de *méthodes de Monte Carlo par Chaînes de Markov* (MCMC). La justification mathématique de l'approximation (1) et le contrôle de l'erreur d'approximation, reposent sur des théorèmes limites pour les chaînes de Markov [13, 8]. Dans ce stage, on considère des lois π définies sur $\mathbb{R}^T$ et ayant une densité (notée encore π) par rapport à la mesure de Lebesgue, de la forme

$$\pi(x) = \exp(-f(x) - g(Ax)) \mathbb{1}_{\mathcal{D}}(x) \quad (2)$$

où f et g sont des fonctions convexes et A est une matrice $T \times T'$; f est une fonction de classe C^1 , Lipschitz sur tout compact, et pouvant posséder un opérateur proximal avec forme explicite; g est une fonction non lisse dont l'opérateur proximal associé a une forme explicite mais celui de $g(A \cdot)$ n'en a pas.

Algorithm 1 une itération d'un HM ciblant π

Require: X_k

Echantillonner $X_{k+1/2} \sim \mu(X_k) + \sqrt{2\gamma}\mathcal{N}_d(0, C)$

Avec probabilité $\alpha(X_k, X_{k+1/2})$,

poser $X_{k+1} = X_{k+1/2}$

et sinon, poser $X_{k+1} = X_k$.

return X_{k+1} .

La méthode MCMC considérée est celle décrite par Algorithm 1 : il s'agit d'un Hastings-Metropolis (HM) avec loi de proposition gaussienne. Le ratio d'acceptation $\alpha(\cdot, \cdot)$ est défini de façon à assurer que la loi π est réversible pour cette transition markovienne; dans le cas où π a un support $\mathcal{D}$, il permet de rejeter tous les points $X_{k+1/2}$ qui sont proposés en dehors de $\mathcal{D}$.

Le choix du la moyenne μ et de la matrice de covariance C se révèlent cruciaux pour l'efficacité des échantillonneurs MCMC (voir par exemple [2, 7]). Différentes stratégies pour le choix de μ ont été proposées, utilisant des informations de gradient ou sous-gradient des fonctions f et g , des opérateurs proximaux associés à f et g , des changements de géométrie lorsque A est inversible pour approcher $\pi(A^{-1} \cdot)$ plutôt que π (voir par exemple [3, 14, 12, 7]). Au-delà d'une comparaison numérique de ces différentes stratégies d'échantillonnage pour une application donnée, il n'existe pas toujours d'études théoriques quantifiant la qualité de convergence des échantillonneurs et permettant, en conséquence, de les comparer.

Objectifs du stage. Le premier objectif du stage est une définition de μ exploitant l'opérateur proximal de f et celui de g mimant le schéma d'optimisation convexe de [6]. Si des schémas numériques et des bornes de convergence existent [11, 10, 4], ces travaux considèrent que tout point proposé est accepté, ce qui ne convient pas lorsque π est de la forme (2) avec $\mathcal{D} \neq \mathbb{R}^T$. Cette contribution est donc à la fois méthodologique, avec la définition d'une procédure d'acceptation-rejet adaptée; et théorique, avec l'étude de convergence des chaînes de Markov en distance de Wasserstein.

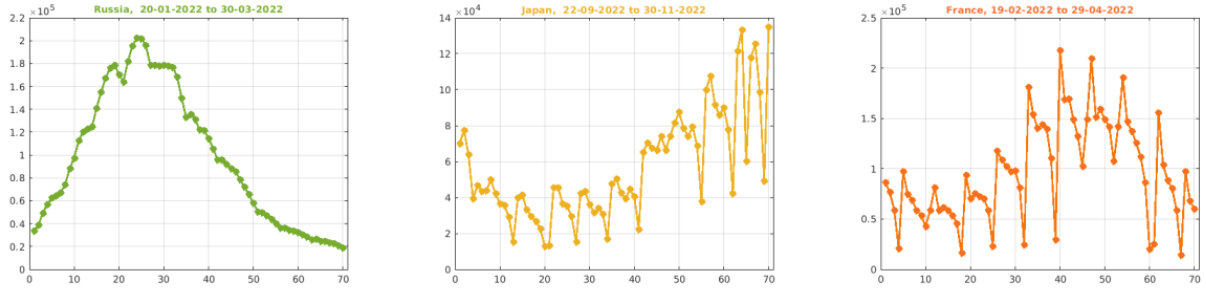


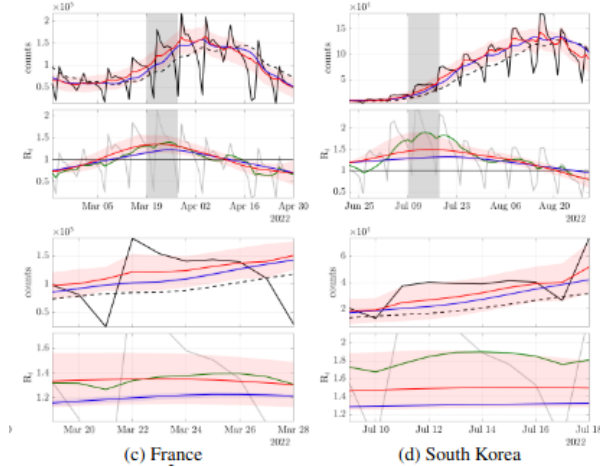
Figure 1: Données de la pandémie Covid19

Le second objectif est applicatif; il est mené dans le cadre du projet ANR OptiMoCSI. La forme de la loi cible (2) est motivée par un problème inverse bayésien en épidémiologie : à partir de données de comptes quotidiens de nouveaux infectés, $\{Z_1, Z_2, \dots, Z_T\}$, on souhaite apprendre l'évolution temporelle d'un *taux de reproduction* d'une épidémie. Le modèle statistique proposé par [1] (voir aussi [9]) est le modèle de Markov caché suivant :

- le processus d'observation $(Z_t)_t$ a pour loi conditionnelle : $\text{Poisson}(R_t \Phi_t + O_t)$ où $\Phi_t := \sum_{u=1}^{\tau} \phi(u) Z_{t-u}$ est une moyenne pondérée des comptes passés sur τ jours;
- le processus caché $((R_t, O_t))_t$ est modélisé comme deux lois de Laplace indépendantes, dont l'une est décentrée

$$R_t \sim 2R_{t-1} - R_{t-2} + 4\text{Lapl}(\lambda_R) \quad O_t \sim \text{Lapl}(\lambda_O).$$

La fonction $t \mapsto R_t$ est le taux de reproduction d'intérêt évoluant en temps, dont la modélisation statistique exprime une condition de variation de la dérivée seconde; $(O_t)_t$ est une famille de variables aléatoires qui modélise les erreurs de comptes et les erreurs de reports des comptes des nouvelles infections;



Graphes extraits de [15]

quant aux observations, elles sont modélisées comme des flux poissonniens dont l'intensité est la somme d'un terme traduisant une contamination de la population par les personnes infectées les τ jours précédents, et d'un terme modélisant un phénomène d'effacement de Poisson dû par exemple à des personnes infectées qui ne se font pas recenser.

Le second objectif du stage est d'appliquer les nouvelles méthodes algorithmiques à la loi a posteriori de R_t sachant les observations $Z_1, \dots, Z_T$, pour $t = T-7, \dots, T$ afin de déterminer une tendance sur l'évolution temporelle du taux de reproduction de la pandémie. Cette tendance sera lue au travers d'une estimation de la moyenne et de quantiles de la loi a posteriori; l'étude de la dispersion de la loi permettra de quantifier l'incertitude.

Pré-requis: Le candidat ou la candidate doit avoir validé des cours de Statistique, et sur la théorie des chaînes de Markov - niveau Master. De plus, avoir validé des cours sur les méthodes d'échantillonnage Monte Carlo ou sur l'optimisation convexe sera aussi apprécié.

Langage de programmation: Les explorations numériques seront développées en Python ou en MATLAB.

Environnement de travail Le ou la stagiaire percevra une gratification de 639,45 € par mois. Le stage se déroulera au sein de l'équipe SARA du LAAS-CNRS, à Toulouse. Le ou la candidate sera également associé-e aux activités de l'équipe Statistiques et Optimisation (SO) de l'Institut de Mathématiques de Toulouse. La participation au séminaire hebdomadaire de l'équipe SO sera fortement encouragée, ainsi qu'aux autres séminaires du site (par exemple SPOT, SOLACE, Machine Learning, etc.).

Candidature: Envoyer un CV et un relevé de notes des deux dernières années, ainsi qu'une lettre de candidature, par email à Gersende Fort et Juliette Chevallier.

Proximal Monte Carlo methods for Pandemic monitoring

Keywords - Monte Carlo sampling, Markov chains, Proximal-based Convex optimization, Applications in epidemiology

Location: LAAS-CNRS Laboratoire d'Analyse et Architecture des Systèmes, Toulouse, France.

Advisors : Gersende Fort (CNRS, LAAS)

email : gersende.fort@cnrs.fr

web : <http://perso.math.univ-toulouse.fr/gfort/>

Juliette Chevallier (INSA Toulouse, IMT)

email : juliette.chevallier@math.univ-toulouse.fr

web : <https://juliette-chevallier.pages.math.cnrs.fr/>

When: 4 to 6 months between mid-february and end of August 2026.

Context – Monte Carlo methods are algorithms that generate a sequence of points $X_k, k \geq 0$ that approximate a probability distribution π [5]. When this distribution is too complex for quantities of interest - such as the mean, variance, or the quantiles of its marginals - to be obtained through explicit formulas, Monte Carlo methods provide a numerical approximation.

$$\int h(x)\pi(dx) \approx \frac{1}{K} \sum_{k=1}^K h(X_k). \quad (3)$$

These points may have Markovian dynamics: the construction of the point X_{k+1} depends on the previous points $(X_{k-r}, \dots, X_{k-1}, X_k)$ and on an independent random mechanism; in this case, we refer to Markov Chain Monte Carlo methods (MCMC). The mathematical justification of the approximation (3), as well as the control of the approximation error, rely on limit theorems for Markov chains [13, 8].

In this internship, we consider distributions π defined on $\mathbb{R}^T$ and having a density (also denoted by π) with respect to the Lebesgue measure, of the form

$$\pi(x) = \exp(-f(x) - g(Ax)) \mathbb{1}_{\mathcal{D}}(x) \quad (4)$$

where f and g are convex functions and A is a $T \times T'$ matrix; f is a C^1 function, Lipschitz on every compact set, and may have a proximal operator in closed form; g is a non-smooth function whose associated proximal operator has a closed-form expression, but the proximal operator of $g(A\cdot)$ does not.

Algorithm 2 One iteration of HM targeting π

Require: X_k

Sample $X_{k+1/2} \sim \mu(X_k) + \sqrt{2\gamma}\mathcal{N}_d(0, C)$

With probability $\alpha(X_k, X_{k+1/2})$,

set $X_{k+1} = X_{k+1/2}$

and otherwise, set $X_{k+1} = X_k$.

return X_{k+1} .

The MCMC method considered is the one described in Algorithm 2: a Hastings–Metropolis (HM) algorithm with a Gaussian proposal distribution. The acceptance–rejection ratio $\alpha(\cdot, \cdot)$ is defined so as to ensure that the distribution π is reversible for this Markov transition. When π has a support $\mathcal{D}$, this mechanism rejects all proposed points $X_{k+1/2}$ that fall outside $\mathcal{D}$.

The choice of the mean μ and the covariance matrix C is crucial for the efficiency of MCMC samplers (see, for example, [2, 7]). Various strategies for selecting μ have been proposed, relying on gradient or subgradient information of the functions f and g , on proximal operators associated with f and g , or geometry-changing transformations when A is invertible, in order to sample from $\pi(A^{-1}\cdot)$ rather than from π (see, for example, [3, 14, 12, 7]). Beyond numerical comparisons of these different sampling strategies for a given application, theoretical studies quantifying the convergence quality of the samplers - and thus allowing them to be compared - do not always exist.

Internship objectives. The first objective of the internship is to define a choice of μ using the proximal operator of f and that of g , mimicking the convex-optimization scheme of [6]. While numerical schemes and convergence bounds exist in the literature [11, 10, 4], these works assume that every proposed point is accepted — which is not appropriate when π is of the form (4) with $\mathcal{D} \neq \mathbb{R}^T$. Therefore, this contribution is both methodological — by defining a suitable acceptance–rejection procedure, and theoretical — by studying the convergence of the Markov chains in Wasserstein distance.

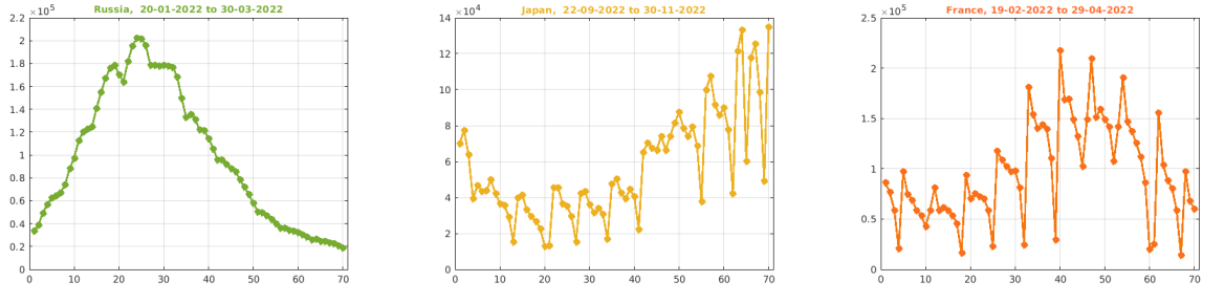


Figure 2: Covid19 data

The second objective is applicative; it is carried out within the framework of the project ANR Opti-MoCSI. The form of the target distribution (4) is motivated by a Bayesian inverse problem in epidemiology: given daily count data of new infections $\{Z_1, Z_2, \dots, Z_T\}$, the goal is to learn the temporal evolution of an epidemic reproduction rate. The statistical model proposed in [1] (see also [9]) is the following hidden Markov model:

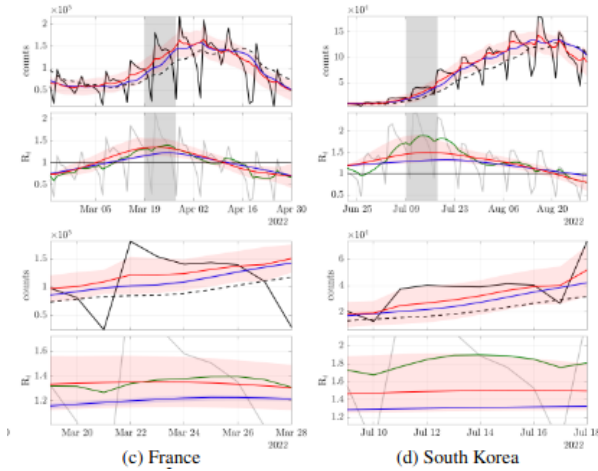
- The observation process $(Z_t)_t$ has conditional distribution $\text{Poisson}(R_t \Phi_t + O_t)$ where $\Phi_t := \sum_{u=1}^{\tau} \phi(u) Z_{t-u}$ is a weighted average of past counts over τ days;
- The hidden process $((R_t, O_t))_t$ is modeled as two independent Laplace distributions, one of which is shifted:

$$R_t \sim 2R_{t-1} - R_{t-2} + 4\text{Lapl}(\lambda_R) \quad O_t \sim \text{Lapl}(\lambda_O).$$

The function $t \mapsto R_t$ is the time-varying reproduction rate of interest, whose statistical modeling expresses a constraint on the variation of its second derivative. The sequence $(O_t)_t$ is a family of random variables modeling counting errors and delays in reporting new infection counts. As for the observations,

they are modeled as Poisson processes whose intensity is the sum of a term representing contamination of the population by individuals infected during the previous τ days, and a term modeling a Poisson thinning phenomenon caused, for example, by infected individuals who are not reported.

The second objective of the internship is to apply the new algorithmic methods to the posterior distribution of R_t given the observations $Z_1, \dots, Z_T$, for $t = T - 7, \dots, T$ in order to determine a trend in the temporal evolution of the pandemic reproduction rate. This trend will be assessed through estimates of the mean and quantiles of the posterior distribution; studying the dispersion of the distribution will allow quantification of the uncertainty.



Plots from [15]

Skills: The candidate must have completed courses in Statistics and Markov chain theory at the Master's level. Additionally, having completed courses on Monte Carlo sampling methods or convex optimization will also be appreciated.

Programming language: The numerical investigations will be developed in Python or in MATLAB.

Work environment. As an intern, you will receive a “gratification” of 639,45€ per month. The research project will be carried out in the SARA team at LAAS, in Toulouse, France. The candidate will also be involved into the activities of the Statistics and Optimization (SO) team of Institut de Mathématiques de Toulouse. Participation to the weekly seminar of the SO team will be strongly encouraged, and to other seminars of the site as well (SPOT, SOLACE, Machine Learning among examples).

Application: Send a CV and a statement of interest by e-mail to Gersende Fort and Juliette Chevallier. Interested candidates can contact the advisors for further information.

References

- [1] P. Abry, J. Chevallier, G. Fort, and B. Pascal. Pandemic intensity estimation from stochastic approximation-based algorithms. In *2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 356–360, 2023.
- [2] 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. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 2196–2200, 2022.
- [3] P. Bianchi, A. Salim, and S. Schechtman. Passty langevin. In *Conference on Machine Learning (CML)*, 2019.
- [4] M. Burger, M.J. Ehrhardt, L. Kuger, and L. Weigand. Analysis of primal-dual langevin algorithms, 2024.
- [5] G. Casella and C.P. Robert. *Monte Carlo Statistical Methods*. Springer, 2004.
- [6] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 2011.
- [7] J. Chevallier and G. Fort. Sampling nonsmooth log-concave densities: A comparative study of primal-dual based proposal distributions. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.
- [8] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer, 2018.
- [9] G. Fort, B. Pascal, P. Abry, and N. Pustelnik. Covid19 reproduction number: Credibility intervals by blockwise proximal monte carlo samplers. *IEEE Transactions on Signal Processing*, 71:888–900, 2023.
- [10] A. Habring, M. Holler, and T. Pock. Subgradient langevin methods for sampling from nonsmooth potentials. *SIAM Journal on Mathematics of Data Science*, 6(4):897–925, 2024.
- [11] T. T.-K. Lau and H. Liu. Bregman proximal Langevin Monte Carlo via Bregman-moreau envelopes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12049–12077, 2022.
- [12] T. T.-K. Lau, H. Liu, and T. Pock. Non-log-concave and nonsmooth sampling via langevin monte carlo algorithms. In Alessandro Benfenati · Federica Porta · Tatiana Alessandra Bubba · Marco Viola, editor, *Advanced Techniques in Optimization for Machine Learning and Imaging*. Springer, 2024.
- [13] S. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.
- [14] D. Narnhofer, A. Habring, M. Holler, and T. Pock. Posterior-variance-based error quantification for inverse problems in imaging. *SIAM Journal on Imaging Sciences*, 17(1):301–333, 2024.
- [15] B. Pascal, P. Abry, J. Chevallier, and G. Fort. A Scaled Poisson Bayesian Model for Viral Epidemic Monitoring. working paper or preprint, September 2025.