Sampling Nonsmooth Log-Concave Densities: A Comparative Study of Primal-Dual Based Proposal Distributions

Gersende Fort

CNRS, France
Laboratoire d'Analyse et d'Architecture des Systèmes

Joint work with J. Chevallier (INSA Toulouse, France; Institut de Mathématiques de Toulouse)

Fundings: ANR-23-CE48-0009, project OptiMoCSI

The problem: Design a new MCMC sampler

when the target distribution has a density π w.r.t. the Lebesgue measure on $\mathbb{R}^d$

- nonsmooth
- log-concave
- restricted to a measurable set $\mathcal{D} \subsetneq \mathbb{R}^d$

More precisely:

$$-\log \pi(\theta) = \begin{cases} f(\theta) + g(\theta) + h(\mathsf{A}\theta) & \text{for} \quad \theta \in \mathcal{D}, \\ +\infty & \text{otherwise,} \end{cases}$$

The set $\mathcal{D}$

· measurable convex subset of $\mathbb{R}^d$
· f, g, h are proper convex functions, with domain including $\mathcal{D}$.

The term f

· f is convex and C^1 on a neighborhood of $\mathcal{D}$

ex. $-f$ is a log-likelihood

The term g

· $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ lower semi-continuous convex function
· $\mathrm{Prox}_{\gamma g}(\cdot)$ has a closed form expression. $\gamma > 0$

Ex. $-g$ is a penalty term.

The term $h(\mathsf{A}\cdot)$

· a $d' \times d$ matrix A
· $h : \mathbb{R}^{d'} \to \mathbb{R} \cup \{+\infty\}$ lower semi-continuous convex function
· $\mathrm{Prox}_{\gamma h}(\cdot)$ has a closed form expression but **not** $\mathrm{Prox}_{\gamma h(\mathsf{A}\cdot)}(\cdot)$. $\gamma > 0$

Ex. $-h(\mathsf{A}\cdot)$ is a penalty term.

Why ?

$-\log \pi(\theta) = f(\theta) + g(\theta) + h(\mathrm{A}\theta)$ **on** $\mathcal{D}$

Applications: nonsmooth convex composite negative log-density in

- **aggregation of estimators** by exponential weighting in PAC-Bayesian learning (Dalalyan and Tsybakov, 2012; Guedj and Alquier, 2013; Luu et al., 2019)
- **Bayesian inverse problems in signal and image processing** (Moulin and Liu, 1999; Dobigeon et al., 2009; Chaâri et al., 2010; Lucka, 2012; Costa et al., 2015; Pereyra, 2016; Chaari et al., 2016; Pascal et al., 2022; Fort et al., 2023)

as few examples.

Not addressed in the literature:

When $\log \pi$ is C^1

- MALA (Roberts and Tweedie, 1996), adapted to $\mathcal{D} \neq \mathbb{R}^d$
- ULA (Durmus and Moulines, 2017; Dalalyan, 2017), not adapted to $\mathcal{D} \neq \mathbb{R}^d$

When $h = 0$

- and $\mathcal{D} = \mathbb{R}^d$: combines MALA/ULA and a smoothing technique based on the Moreau-Yoshida envelope (Pereyra, 2016; Durmus et al., 2018; Pereyra et al., 2020; Luu et al., 2021; Durmus et al., 2022; Lau and Liu, 2022).
- and $\mathcal{D} \subsetneq \mathbb{R}^d$, Langevin-based methods + projections/reflections: for specific topologies of $\mathcal{D}$ (Bubeck et al., 2018; Melidonis et al., 2023).

With $h \neq 0$ but $\mathcal{D} = \mathbb{R}^d$, ULA-based methods (Luu et al., 2021).

In our setting: our previous contribution (Fort et al., 2023).

Hastings-Metropolis combined with first-order optimization methods

$-\ln \pi(\theta) = f(\theta) + g(\theta) + h(\mathsf{A}\theta)$ **on** $\mathcal{D}$

- A Hastings-Metropolis step to deal with the domain $\mathcal{D}$

$$\theta_{t+1/2} = \mu_\gamma(\theta_t) + \mathcal{N}_d\left(0, \mathsf{C}\right)$$
$$\theta_{t+1} = \text{Acceptance-Rejection} \quad \text{step}\left(\theta_t, \theta_{t+1/2}\right).$$

Propose without constraints. Reject any candidate which is not in $\mathcal{D}$.

- A drift term $\mu_\gamma(\theta)$ based on first order informations on $\log \pi$

- The rate of ergodicity of a Markov chain is related to how fast, starting from a small set, the chain returns back to this small set (Meyn and Tweedie, 1993, Theorem 15.0.1.)
- Find drift terms that can push the chain towards the modes of π when the chain is visiting the tails of π.

> Choose $\mu_\gamma(\theta)$ as one step of a convex optimization procedure for minimizing $-\log \pi$

- Characterization of the optimum $\theta_\star$ see e.g. (Bauschke and Combettes, 2019, Theorem 16.3)

$$0 = \nabla f(\theta_\star) + u_\star + \mathsf{A}^\top s_\star \qquad u_\star \in \partial g(\theta_\star) \qquad s_\star \in (\partial h)(\mathsf{A}\theta_\star)$$

Find μ_γ s.t. $\theta_\star = \mu_\gamma(\theta_\star)$.

Two methods

The optimum $\theta_\star$ solves $0 = \nabla f(\theta_\star) + u_\star + \mathsf{A}^\top s_\star \qquad u_\star \in \partial g(\theta_\star),\ s_\star \in \partial h(\mathsf{A}\theta_\star)$

- Full sub-gradient

$$\mu_\gamma(\theta) = \theta - \gamma\left(\nabla f(\theta) + G(\theta) + \mathsf{A}^\top H(\mathsf{A}\theta)\right) \qquad \begin{aligned} G(\theta) &\in \partial g(\theta) \\ H(\tau) &\in \partial h(\tau) \end{aligned}$$

- Proximal sub-gradient

Prop 1. $-\nabla f(\theta_\star) - \mathsf{A}^\top s_\star = u_\star$

Prop 2. $u_\star \in \partial g(\theta_\star)$ iff $\theta_\star = \mathrm{Prox}_{\gamma g}(\theta_\star + \gamma u_\star)$

$$\mu_\gamma(\theta) = \mathrm{Prox}_{\gamma g}\left(\theta - \gamma\nabla f(\theta) - \gamma\mathsf{A}^\top H(\mathsf{A}\theta)\right) \qquad H(\tau) \in \partial h(\tau)$$

Proposal mechanism: $\mu_\gamma(\theta) + \sqrt{2\gamma}\,\mathcal{N}_d\,(0; \mathrm{I})$.

Two other methods when A is invertible

The optimum $\theta_\star$ solves $0 = \nabla f(\theta_\star) + u_\star + \mathrm{A}^\top s_\star \qquad u_\star \in \partial g(\theta_\star),\ s_\star \in \partial h(\mathrm{A}\theta_\star)$

Lemma: $(\theta_t)_t$ is a Markov chain with invariant distribution $\pi(\cdot)$ iff $(\mathrm{A}\theta_t)_t$ is a Markov chain with invariant distribution $\propto \pi(\mathrm{A}^{-1}\cdot)$.

- inv Full sub-gradient

(Fort et al., 2023)

$$\mu_\gamma(\theta) = \theta - \gamma \mathrm{A}^{-1}\mathrm{A}^{-\top}\left(\nabla f(\theta) + G(\theta) + \mathrm{A}^\top H(\mathrm{A}\theta)\right)$$

$$G(\theta) \in \partial g(\theta)$$
$$H(\tau) \in \partial h(\tau)$$

- inv Sub-gradient proximal

Prop 1. $-\mathrm{A}^{-\top}\nabla f(\theta_\star) - \mathrm{A}^{-\top} u_\star = s_\star$

Prop 2. $s_\star \in \partial h(\mathrm{A}\theta_\star)$ iff $\mathrm{A}\theta_\star = \mathrm{Prox}_{\gamma h}(\mathrm{A}\theta_\star + \gamma \mathrm{A} s_\star)$

$$\mu_\gamma(\theta) = \mathrm{A}^{-1}\mathrm{Prox}_{\gamma h}\left(\mathrm{A}\theta - \gamma \mathrm{A}^{-\top}\nabla f(\theta) - \gamma \mathrm{A}^{-\top} G(\theta)\right) \qquad G(\tau) \in \partial g(\tau)$$

Proposal mechanism: $\mu_\gamma(\theta) + \sqrt{2\gamma}\,\mathcal{N}_d\left(0; \mathrm{A}^{-1}\mathrm{A}^{-\top}\right)$.

Applications: Bayesian inference for count data in epidemiology

$-\ln \pi(\theta) = f(\theta) + g(\theta) + h(\mathsf{A}\theta)$ **on** $\mathcal{D}$

▶ The HMM model (Fort et al., 2023, Section II-C)

· Count data: $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$

· Set $\Phi_t := \sum_{u=1}^{\tau} \phi_u \mathsf{Z}_{t-u}$

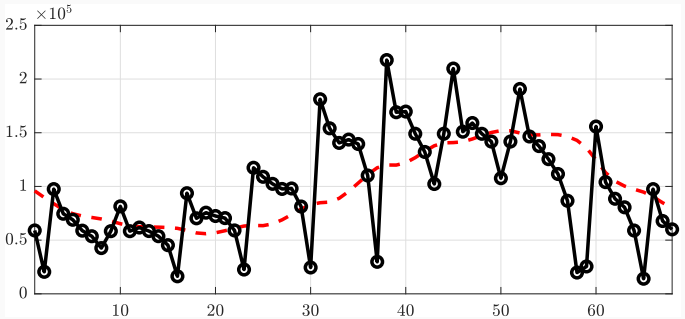
· $\mathsf{Z}_t | \mathsf{R}_t, \mathsf{O}_t, \mathrm{past}_{t-1} \sim \mathcal{P}(\mathsf{R}_t \Phi_t + \mathsf{O}_t)$

· $\mathsf{R}_t | \mathrm{past}_{t-1} \sim 2\mathsf{R}_{t-1} - \mathsf{R}_{t-2} + \mathrm{Lapl}(\lambda_{\mathsf{R}}/4)$

· $\mathsf{O}_t | \mathrm{past}_{t-1} \sim \mathrm{Lapl}(\lambda_{\mathsf{O}})$

· $(\mathsf{R}_0, \mathsf{R}_{-1}, \mathsf{Z}_0, \cdots, \mathsf{Z}_{1-\tau}) \in \mathrm{past}_0$

Explore: $(\mathsf{R}_1, \cdots, \mathsf{R}_T, \mathsf{O}_1, \cdots, \mathsf{O}_T) \in \mathbb{R}^{2T}$



The daily new infection counts $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$ (black curve) and the averaged past counts $\Phi_1, \cdots, \Phi_T$ (dashed red curve). Covid-19 data, France, from 2022/02/20 to 2022/04/28

▶ The a posteriori distribution of $\theta := (\mathsf{R}_1, \cdots, \mathsf{R}_T, \bar{\mathsf{O}}_1, \cdots, \bar{\mathsf{O}}_T) \in \mathbb{R}^{2T}$

For numerical considerations: $\mathsf{O}_t = \Phi_t \bar{\mathsf{O}}_t$

· with $f = 0$ and g given by or with $g = 0$ and f given by

$$\theta \mapsto \sum_{t=1}^{T} \left(\Phi_t (\mathsf{R}_t + \bar{\mathsf{O}}_t) - \mathsf{Z}_t \ln(\mathsf{R}_t + \bar{\mathsf{O}}_t) \right)$$

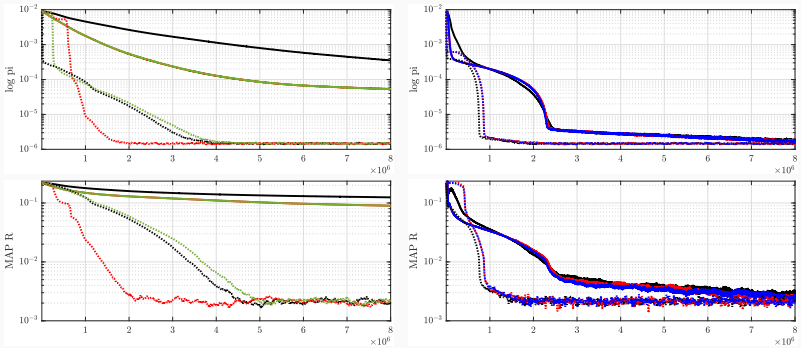
· with h and A s.t. $h(\mathsf{A}\cdot)$ given by

$\theta \mapsto \lambda_{\mathsf{R}} \|\mathsf{D}\mathbf{R} + \delta\|_1 + \lambda_{\mathsf{O}} \|\mathbf{\Phi}\bar{\mathbf{O}}\|_1$, A block diagonal, with blocks D and $\lambda_{\mathsf{O}}/\lambda_{\mathsf{R}}\Phi$

· and $\mathcal{D}$ is

$$\bigcap_{t:\mathsf{Z}_t>0} \{(\mathbf{R}, \bar{\mathbf{O}}) : \mathsf{R}_t + \bar{\mathsf{O}}_t > 0\} \cap \bigcap_{t:\mathsf{Z}_t=0} \{(\mathbf{R}, \bar{\mathbf{O}}) : \mathsf{R}_t + \bar{\mathsf{O}}_t \geq 0\}$$

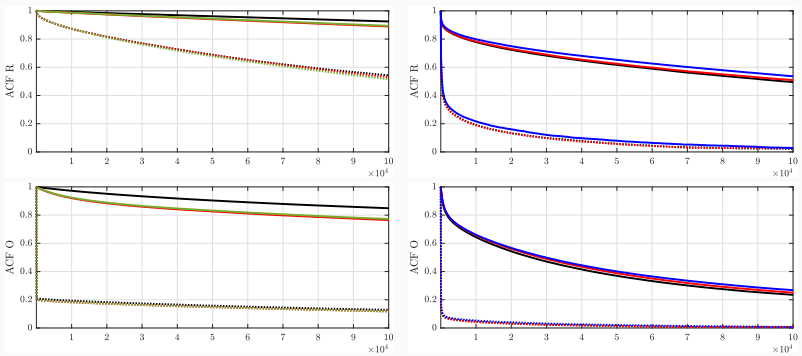
Comparison of the proposal mechanisms



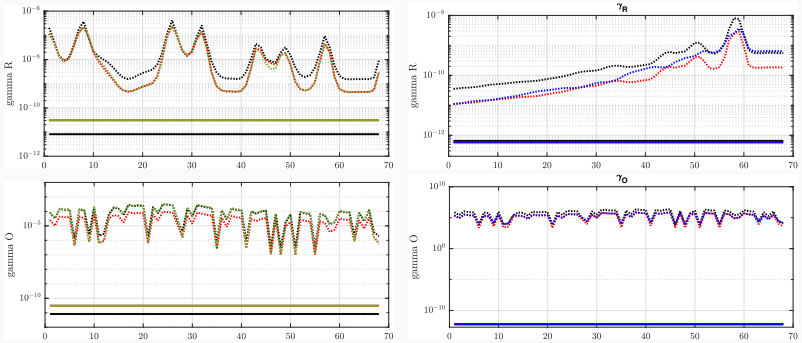
(top) Normalized distance to the optimum of $\log \pi$.
(bottom) Normalized distance to the MAP **R**

· Natural geometry: RW, FSG, Prox-SG
· After a change of geometry (use A^{-1}) inv-RW, inv-FSG, SG-Prox

· `Full update` are in solid lines
· `one-at-a-time Gibbs` samplers are in dotted lines.



mean abs(ACF) of the R_t components (top) and the $\bar{O}_t$ components (bottom)



the step size γ at the end of the burn-in period

Conclusions: (i) Change of geometry → covariance structure; (ii) block / one-at-a-time updates; (iii) FSG strategies

Conclusions and perspectives

- We compared different strategies for the definition of the drift term in a HM sampler with a Gaussian proposal, that uses first-order information on $\log \pi$.
- What about the use of `fully proximal` drift terms by adapting the primal-dual optimization method `PD3O` (Ming, 2018) ?
- MALA is known to be poor for heavy-tailed distributions, except when combined with preconditioning strategies (Fort and Roberts, 2005): what about primal-dual based methods with state-dependent preconditioners ?
- Application to inference of the reproduction number :
  - what about the design parameters (λ_R, λ_O) ?
  - what about the estimation of the R_t's when using such MCMC samplers ?

  see the talk* by P. Abry (CNRS, France) in ICASSP 2025.

*Session: Bayesian Signal Processing I; title of the talk "Hierarchical Bayesian Estimation of COVID-19 Reproduction Number "

References i

Bauschke, H. and Combettes, P.-L. (2019). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer.

Bubeck, S., Eldan, R., and Lehec, J. (2018). Sampling from a Log-Concave Distribution with Projected Langevin Monte Carlo. *Discrete Comput. Geom.*, 59:757–783.

Chaari, L., Tourneret, J.-Y., Chaux, C., and Batatia, H. (2016). A Hamiltonian Monte Carlo Method for Non-Smooth Energy Sampling. *IEEE Trans. Signal Process.*, 64(21):5585–5594.

Chaâri, L., Pesquet, J.-C., Tourneret, J.-Y., Ciuciu, P., and Benazza-Benyahia, A. (2010). A hierarchical Bayesian model for frame representation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4086–4089.

Costa, F., Batatia, H., Chaari, L., and Tourneret, J.-Y. (2015). Sparse EEG Source Localization Using Bernoulli Laplacian Priors. *IEEE Trans. Biomed. Eng.*, 62(12):2888–2898.

Dalalyan, A. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B Methodol.*, 79(3):651–676.

Dalalyan, A. and Tsybakov, A. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. Syst. Sci.*, 78(5):1423–1443.

Dobigeon, N., Hero, A. O., and Tourneret, J.-Y. (2009). Hierarchical Bayesian Sparse Image Reconstruction With Application to MRFM. *IEEE Trans. Image Process.*, 18(9):2059–2070.

Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551 – 1587.

Durmus, A., Moulines, E., and Pereyra, M. (2018). Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM J. Imaging Sci.*, 11(1):473–506.

Durmus, A., Moulines, E., and Pereyra, M. (2022). A Proximal Markov Chain Monte Carlo Method for Bayesian Inference in Imaging Inverse Problems: When Langevin Meets Moreau. *SIAM Review*, 64(4):991–1028.

Fort, G., Pascal, B., Abry, P., and Pustelnik, N. (2023). Covid19 reproduction number: Credibility intervals by blockwise proximal monte carlo samplers. *IEEE Trans. Signal Process.*, 71:888–900.

Fort, G. and Roberts, G. O. (2005). Subgeometric ergodicity of strong Markov processes. *Ann. Appl. Probab.*, 15(2):1565 – 1589.

References ii

Guedj, B. and Alquier, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *ElectroN. J. Stat.*, 7(none):264 – 291.

Lau, T. T.-K. and Liu, H. (2022). Bregman proximal Langevin Monte Carlo via Bregman-moreau envelopes. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12049–12077.

Lucka, F. (2012). Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors. *Inverse Probl.*, 28(12):125012.

Luu, T., Fadili, J., and Chesneau, C. (2019). Pac-bayesian risk bounds for group-analysis sparse regression by exponential weighting. *J. Multivar. Anal.*, 171:209–233.

Luu, T., Fadili, J., and Chesneau, C. (2021). Sampling from Non-smooth Distributions Through Langevin Diffusion. *Methodology and Computing in Applied Probability*, 23:1173–1201.

Melidonis, S., Dobson, P., Altmann, Y., Pereyra, M., and Zygalakis, K. (2023). Efficient bayesian computation for low-photon imaging problems. *SIAM J. Imaging Sci.*, 16(3):1195–1234.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

Ming, Y. (2018). A New Primal–Dual Algorithm for Minimizing the Sum of Three Functions with a Linear Operator. *J. Sci. Comput.*, 76:1698–1717.

Moulin, P. and Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Trans. Inf. Theory*, 45(3):909–919.

Pascal, B., Abry, P., Pustelnik, N., G, R. S., Gribonval, R., and Flandrin, P. (2022). Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low quality data. *IEEE Trans. Signal Process.*, 70:2859–2868.

Pereyra, M. (2016). Proximal Markov chain Monte Carlo algorithms. *Stat. Comput.*, 26:745–760.

Pereyra, M., Mieles, L. V., and Zygalakis, K. C. (2020). Accelerating Proximal Markov Chain Monte Carlo by Using an Explicit Stabilized Method. *SIAM J. Imaging Sci.*, 13(2):905–935.

Roberts, G. and Tweedie, R. (1996). Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363.