



**HAL**  
open science

# Sampling from non-smooth distribution through Langevin diffusion

Duy Tung Luu, Jalal M. Fadili, Christophe Chesneau

► **To cite this version:**

Duy Tung Luu, Jalal M. Fadili, Christophe Chesneau. Sampling from non-smooth distribution through Langevin diffusion. ORASIS 2017, GREYC, Jun 2017, Colleville-sur-Mer, France. hal-01866621

**HAL Id: hal-01866621**

**<https://hal.science/hal-01866621>**

Submitted on 3 Sep 2018

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

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

# Agrégation à poids exponentiels : Algorithmes d'échantillonnage

Luu Duy Tung<sup>1</sup>

Jalal Fadili<sup>1</sup>

Christophe Chesneau<sup>2</sup>

<sup>1</sup> Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, France

<sup>2</sup> Normandie Univ, UNICAEN, CNRS, LMNO, France

duy-tung.luu@ensicaen.fr

jalal.fadili@ensicaen.fr

christophe.chesneau@unicaen.fr

## Résumé

Nous proposons dans cet article des algorithmes d'échantillonnage de distributions dont la densité est ni lisse ni log-concave. Nos algorithmes sont basés sur la diffusion de Langevin de la densité lissée par la régularisation de Moreau-Yosida. Ces résultats sont ensuite appliqués pour établir des agrégats à poids exponentiels dans un contexte de grande dimension.

## Mots Clef

Diffusion de Langevin, régularisation de Moreau-Yosida, agrégation à poids exponentiels.

## Abstract

In this paper, we propose algorithms for sampling from the distributions whose density is non-smoothed nor log-concave. Our algorithms are based on the Langevin diffusion on the regularized counterpart of density by the Moreau-Yosida regularization. These results are then applied to compute the exponentially weighted aggregates for high dimensional regression.

## Keywords

Langevin diffusion, Moreau-Yosida smoothing, exponentially weighted aggregation.

## 1 Introduction

Consider the following linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi} \quad (1)$$

where  $\mathbf{y} \in \mathbb{R}^n$  is the response vector,  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is a deterministic design matrix, and  $\boldsymbol{\xi}$  are errors. The objective is to estimate the vector  $\boldsymbol{\theta}_0 \in \mathbb{R}^p$  from the observations in  $\mathbf{y}$ . Generally, the problem (1) is either under-determined or determined (i.e.,  $p \leq n$ ), but  $\mathbf{X}$  is ill-conditioned, and then (1) becomes ill-posed. However,  $\boldsymbol{\theta}_0$  generally verifies some notions of low-complexity. Namely, it has either a simple structure or a small intrinsic dimension. One can impose the notion of low-complexity on the estimators by considering a prior promoting it.

**Exponential weighted aggregation (EWA)** EWA consists to calculate the following expectation

$$\hat{\boldsymbol{\theta}}_n^{\text{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta} \hat{\mu}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad \hat{\mu}(\boldsymbol{\theta}) \propto \exp(-V(\boldsymbol{\theta})/\beta), \quad (2)$$

where  $\beta > 0$  is called temperature parameter and

$$V(\boldsymbol{\theta}) \stackrel{\text{def}}{=} F(\mathbf{X}\boldsymbol{\theta}, \mathbf{y}) + W_\lambda \circ \mathbf{D}^\top(\boldsymbol{\theta}),$$

where  $F : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  is a general loss function assumed to be differentiable,  $W_\lambda : \mathbb{R}^q \rightarrow \mathbb{R} \cup \{+\infty\}$  is a regularizing penalty depending on a parameter  $\lambda > 0$ , and  $\mathbf{D} \in \mathbb{R}^{p \times q}$  a analysis operator.  $W_\lambda$  promotes some specific notion of low-complexity.

**Langevin diffusion** The computation of  $\hat{\boldsymbol{\theta}}_n^{\text{EWA}}$  corresponds to an integration problem which becomes very involved to solve analytically, or even numerically in high-dimension. A classical approach is to approximate it using a Markov chain Monte-Carlo (MCMC) method which consists in sampling from  $\mu$  by constructing a Markov chain via the Langevin diffusion process, and to compute sample path averages based on the output of the Markov chain. A Langevin diffusion  $\mathbf{L}$  in  $\mathbb{R}^p$ ,  $p \geq 1$  is a homogeneous Markov process defined by the stochastic differential equation (SDE)

$$d\mathbf{L}(t) = \frac{1}{2}\boldsymbol{\rho}(\mathbf{L}(t))dt + d\mathbf{W}(t), \quad t > 0, \quad \mathbf{L}(0) = \mathbf{l}_0, \quad (3)$$

where  $\boldsymbol{\rho} = \nabla \log \mu$ ,  $\mu$  is everywhere non-zero and suitably smooth target density function on  $\mathbb{R}^p$ ,  $\mathbf{W}$  is a  $p$ -dimensional Brownian process and  $\mathbf{l}_0 \in \mathbb{R}^p$  is the initial value. Under mild assumptions, the SDE (3) has a unique strong solution and,  $\mathbf{L}(t)$  has a stationary distribution with density  $\mu$ . This opens the door to approximating integrals  $\int_{\mathbb{R}^p} f(\boldsymbol{\theta})\mu(\boldsymbol{\theta})d\boldsymbol{\theta}$ , where  $f : \mathbb{R}^p \rightarrow \mathbb{R}$ , by the average value of a Langevin diffusion, i.e.,  $\frac{1}{T} \int_0^T f(\mathbf{L}(t))dt$  for a large enough  $T$ . In practice, we cannot follow exactly the dynamic defined by the SDE (3). Instead, we must discretize it by the forward (Euler) scheme, which reads

$$\mathbf{L}_{k+1} = \mathbf{L}_k + \frac{\delta}{2}\boldsymbol{\rho}(\mathbf{L}_k) + \sqrt{\delta}\mathbf{Z}_k, \quad t > 0, \quad \mathbf{L}_0 = \mathbf{l}_0,$$

where  $\delta > 0$  is a sufficiently small constant discretization step-size and  $\{\mathbf{Z}_k\}_k$  are iid  $\sim \mathcal{N}(0, \mathbf{I}_p)$ . The average value  $\frac{1}{T} \int_0^T \mathbf{L}(t)dt$  can then be naturally approximated via the Riemann sum  $\delta/T \sum_{k=0}^{\lfloor T/\delta \rfloor - 1} \mathbf{L}_k$  where  $\lfloor T/\delta \rfloor$  denotes the integer part of  $T/\delta$ . For a complete review about sampling by Langevin diffusion from smooth and log-concave densities, we refer the studies in [1]. To cope with non-smooth

densities, several works have proposed to replace  $\log \mu$  with a smoothed version (typically involving the Moreau-Yosida regularization) [2, 5, 3, 4].

## 2 Algorithm and guarantees

Our main contribution is to enlarge the family of  $\mu$  covered in [2, 5, 3, 4] by relaxing the underlying conditions. Namely, in our framework,  $\mu$  is structured as  $\hat{\mu}$  with  $W_\lambda$  is not necessarily differentiable nor convex. Let  $F_\beta = F(\mathbf{X} \cdot, \mathbf{y})/\beta$  and  $W_{\beta,\lambda} = W_\lambda/\beta$ . To apply the Langevin Monte-Carlo approach, we regularize  $W_{\beta,\lambda}$  by a Moreau envelope defined as

$$\gamma W_{\beta,\lambda}(\mathbf{u}) \stackrel{\text{def}}{=} \inf_{\mathbf{w} \in \mathbb{R}^q} \frac{\|\mathbf{w} - \mathbf{u}\|_2^2}{2\gamma} + W_{\beta,\lambda}(\mathbf{w}), \quad \gamma > 0.$$

Define also the corresponding proximal mapping as

$$\text{prox}_{\gamma W_{\beta,\lambda}}(\mathbf{u}) \stackrel{\text{def}}{=} \text{Argmin}_{\mathbf{w} \in \mathbb{R}^q} \frac{\|\mathbf{w} - \mathbf{u}\|_2^2}{2\gamma} + W_{\beta,\lambda}(\mathbf{w}), \quad \gamma > 0.$$

To establish the algorithm, let us state some assumptions.

- (H.1)  $W_{\beta,\lambda}$  is proper, lsc and bounded from below.
- (H.2)  $\text{prox}_{\gamma W_{\beta,\lambda}}$  is single valued.
- (H.3)  $\text{prox}_{\gamma W_{\beta,\lambda}}$  is locally Lipschitz continuous.
- (H.4)  $\exists K_1 > 0, \forall \boldsymbol{\theta} \in \mathbb{R}^p, \langle \mathbf{D}^\top \boldsymbol{\theta}, \text{prox}_{\gamma W_{\beta,\lambda}}(\mathbf{D}^\top \boldsymbol{\theta}) \rangle \leq K(1 + \|\boldsymbol{\theta}\|_2^2)$ .
- (H.5)  $\exists K_2 > 0, \forall \boldsymbol{\theta} \in \mathbb{R}^p, \langle \boldsymbol{\theta}, \nabla F_\beta(\boldsymbol{\theta}) \rangle \leq K_2(1 + \|\boldsymbol{\theta}\|_2^2)$ .

A large family of  $W_{\beta,\lambda}$  satisfies (H.1)-(H.3). Indeed, one can show that the functions called prox-regular (and a fortiori convex) functions verify these assumptions. The following proposition ensures differentiability of  $W_{\beta,\lambda}$  and expresses the gradient  $\nabla^\gamma W_{\beta,\lambda}$  through  $\text{prox}_{\gamma W_{\beta,\lambda}}$ .

**Proposition 2.1.** *Assume that (H.1)-(H.2) hold. Then  $\gamma W_{\beta,\lambda} \in C^1(\mathbb{R}^q)$  with  $\nabla^\gamma W_{\beta,\lambda} = \frac{1}{\gamma}(\mathbf{I}_q - \text{prox}_{\gamma W_{\beta,\lambda}})$ .*

Consider the Langevin diffusion  $\mathbf{L} \in \mathbb{R}^p$  defined by the following SDE

$$d\mathbf{L}(t) = -\frac{1}{2}\nabla\left(F_\beta + (\gamma W_{\beta,\lambda}) \circ \mathbf{D}^\top\right)(\mathbf{L}(t))dt + d\mathbf{W}(t), \quad (4)$$

when  $t > 0$  and  $\mathbf{L}(0) = \mathbf{l}_0$ . Here  $\mathbf{W}$  is a  $p$ -dimensional Brownian process and  $\mathbf{l}_0 \in \mathbb{R}^p$  is the initial value.

**Proposition 2.2.** *Assume that (H.1)-(H.5) hold. For every initial point  $\mathbf{L}(0)$  such that  $\mathbb{E}[\|\mathbf{L}(0)\|_2^2] < \infty$ , SDE (4) has a unique solution which is strongly Markovian, non-explosive and admits an unique invariant measure  $\hat{\mu}_\gamma \propto \exp\left(-\left(F_\beta(\boldsymbol{\theta}) + (\gamma W_{\beta,\lambda}) \circ \mathbf{D}^\top(\boldsymbol{\theta})\right)\right)$ .*

The following proposition answers the natural question on the behaviour of  $\hat{\mu}_\gamma - \hat{\mu}$  as a function of  $\gamma$ .

**Proposition 2.3.** *Assume that (H.1) hold. Then  $\hat{\mu}_\gamma$  converges to  $\hat{\mu}$  in total variation as  $\gamma \rightarrow 0$ .*

Inserting the identities of Lemma 2.1 into (4), we get

$$d\mathbf{L}(t) = \mathcal{A}(\mathbf{L}(t))dt + d\mathbf{W}(t), \quad \mathbf{L}(0) = \mathbf{l}_0, \quad t > 0. \quad (5)$$

where  $\mathcal{A} = -\frac{1}{2}\left(\nabla F_\beta + \gamma^{-1}\mathbf{D}\left(\mathbf{I}_q - \text{prox}_{\gamma W_{\beta,\lambda}}\right) \circ \mathbf{D}^\top\right)$ . Consider now the forward Euler discretization of (5) with step-size  $\delta > 0$ , which can be rearranged as

$$\mathbf{L}_{k+1} = \mathbf{L}_k + \delta\mathcal{A}(\mathbf{L}_k) + \sqrt{\delta}\mathbf{Z}_k, \quad t > 0, \quad \mathbf{L}_0 = \mathbf{l}_0. \quad (6)$$

From (6), an Euler approximate solution is defined as

$$\mathbf{L}^\delta(t) \stackrel{\text{def}}{=} \mathbf{L}_0 + \int_0^t \mathcal{A}(\bar{\mathbf{L}}(s))ds + \int_0^t d\mathbf{W}(s)ds,$$

where  $\bar{\mathbf{L}}(t) = \mathbf{L}_k$  for  $t \in [k\delta, (k+1)\delta[$ . Observe that  $\mathbf{L}^\delta(k\delta) = \bar{\mathbf{L}}(k\delta) = \mathbf{L}_k$ , hence  $\mathbf{L}^\delta(t)$  and  $\bar{\mathbf{L}}(t)$  are continuous-time extensions to the discrete-time chain  $\{\mathbf{L}_k\}_k$ . Mean square convergence of the pathwise approximation (6) and of its first-order moment is described below.

**Theorem 2.1.** *Assume that (H.1)-(H.5) hold, and  $\mathbb{E}[\|\mathbf{L}(0)\|_2^2] < \infty$  for any  $p \geq 2$ . Then*

$$\|\mathbb{E}[\mathbf{L}^\delta(T)] - \mathbb{E}[\mathbf{L}(T)]\|_2 \leq \mathbb{E}\left[\sup_{0 \leq t \leq T} \|\mathbf{L}^\delta(t) - \mathbf{L}(t)\|_2\right] \xrightarrow{\delta \rightarrow 0} 0.$$

Our algorithm has been applied in several numerical problems. Figure 1 shows an application in Inpainting using EWA with SCAD and  $\ell_{1,2}$  penalties.



FIGURE 1 – (a) : Masked image (b) : Inpainting with EWA -  $\ell_{1,2}$ . (c) Inpainting with EWA - SCAD.

## Références

- [1] A. S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. to appear in JRSS B 1412.7392, arXiv, 2014.
- [2] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *J. Comput. Syst. Sci.*, 78(5):1423–1443, Sept. 2012.
- [3] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. Preprint hal-01176132, July 2015.
- [4] A. Durmus, E. Moulines, and M. Pereyra. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin. hal-01267115, 2016.
- [5] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4), 2016.