Agrégation à poids exponentiels : Algorithmes d’échantillonnage

Luu Duy Tung1  Jalal Fadili1  Christophe Chesneau2
1 Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, France
2 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, exponential weighted aggregation.

1 Introduction
Consider the following linear regression
\[ y = X\theta_0 + \xi \] (1)
where \( y \in \mathbb{R}^n \) is the response vector, \( X \in \mathbb{R}^{n \times p} \) is a deterministic design matrix, and \( \xi \) are errors. The objective is to estimate the vector \( \theta_0 \in \mathbb{R}^p \) from the observations in \( y \). Generally, the problem (1) is either under-determined or determined (i.e., \( p \leq n \)), but \( X \) is ill-conditioned, and then (1) becomes ill-posed. However, \( \theta_0 \) 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
\[ \hat{\theta}_n^{\text{EWA}} = \int_{\mathbb{R}^p} \theta \tilde{\mu}(\theta) d\theta, \quad \tilde{\mu}(\theta) \propto \exp(-V(\theta)/\beta), \] (2)
where \( \beta > 0 \) is called temperature parameter and
\[ V(\theta) = F(X\theta, y) + W_\lambda \circ D^T(\theta), \]
where \( F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) is a general loss function assumed to be differentiable, \( W_\lambda : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) is a regularizing penalty depending on a parameter \( \lambda > 0 \), and \( D \in \mathbb{R}^{p \times q} \) is an analysis operator. \( W_\lambda \) promotes some specific notion of low-complexity.

Langevin diffusion The computation of \( \hat{\theta}_n^{\text{EWA}} \) 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 \( L \) in \( \mathbb{R}^p, p \geq 1 \) is a homogeneous Markov process defined by the stochastic differential equation (SDE)
\[ dL(t) = \frac{1}{2}\rho(L(t))dt + dW(t), \quad t > 0, \quad L(0) = l_0, \] (3)
where \( \rho = \nabla \log \mu, \mu \) is everywhere non-zero and suitably smooth target density function on \( \mathbb{R}^p \), \( W \) is a \( p \)-dimensional Brownian process and \( l_0 \in \mathbb{R}^p \) is the initial value. Under mild assumptions, the SDE (3) has a unique strong solution and, \( L(t) \) has a stationary distribution with density \( \mu \). This opens the door to approximating integrals \( \int_{\mathbb{R}^p} f(\theta)\mu(\theta)d\theta, \) where \( f : \mathbb{R}^p \to \mathbb{R} \), by the average value of a Langevin diffusion, i.e., \( \frac{1}{\delta} \int_0^T f(L(t))dt \) 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
\[ L_{k+1} = L_k + \delta^2 \rho(L_k) + \sqrt{\delta}Z_k, \quad t > 0, \quad L_0 = l_0, \]
where \( \delta > 0 \) is a sufficiently small constant discretization step-size and \( \{ Z_k \}_k \) are iid \( \mathcal{N}(0, I_p) \). The average value \( \frac{1}{\delta} \int_0^T L(t)dt \) can then be naturally approximated via the Riemann sum \( \delta/T \sum_{k=0}^{T/\delta} L_k \) where \( [T/\delta] \) 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 \( \lambda \) is not necessarily differentiable nor convex. Let \( F_\beta = F(X, y)/\beta \) 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

\[
\gamma W_{\beta, \lambda}(u) \overset{\text{def}}{=} \inf_{w \in \mathbb{R}^p} \frac{\|w - u\|^2}{2\gamma} + W_{\beta, \lambda}(w), \quad \gamma > 0.
\]

Define also the corresponding proximal mapping as

\[
\text{prox}_\gamma W_{\beta, \lambda}(u) \overset{\text{def}}{=} \text{Argmin}_{w \in \mathbb{R}^p} \frac{\|w - u\|^2}{2\gamma} + W_{\beta, \lambda}(w), \quad \gamma > 0.
\]

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 \theta \in \mathbb{R}^p, \langle D^\top \theta, \text{prox}_\gamma W_{\beta, \lambda}(D^\top \theta) \rangle \leq K_1(1 + \|\theta\|^2_2).

(H.5) \( \exists K_2 > 0, \forall \theta \in \mathbb{R}^p, \langle \theta, \nabla F_\beta(\theta) \rangle \leq K_2(1 + \|\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 \( \nabla W_{\beta, \lambda} \) through \( \text{prox}_{\gamma W_{\beta, \lambda}} \).

**Proposition 2.1.** Assume that (H.1)-(H.2) hold. Then \( \gamma W_{\beta, \lambda} \in C^1(\mathbb{R}^p) \) with \( \nabla W_{\beta, \lambda} = \frac{1}{\gamma} \left( I_q - \text{prox}_\gamma W_{\beta, \lambda} \right) \).

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

\[
dL(t) = -\frac{1}{2} \nabla \left( F_\beta + (\gamma W_{\beta, \lambda}) \circ D^\top \right)(L(t)) dt + dW(t), \quad t > 0.
\]

when \( t > 0 \) and \( L(0) = l_0 \). Here \( W \) is a p-dimensional Brownian process and \( l_0 \in \mathbb{R}^p \) is the initial value.

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

The following proposition answers the natural question on the behaviour of \( \hat{\mu}_\gamma \) 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 \( \gamma \to 0 \).

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

\[
dL(t) = A(L(t)) dt + dW(t), \quad L(0) = l_0, \quad t > 0.
\]

where \( A = -\frac{1}{2} \left( \nabla F_\beta + \gamma^{-1} D \left( I_q - \text{prox}_\gamma W_{\beta, \lambda} \circ D^\top \right) \right) \).

Consider now the forward Euler discretization of (5) with step-size \( \delta > 0 \), which can be rearranged as

\[
L_{k+1} = L_k + \delta A(L_k) + \sqrt{\delta} Z_k, \quad t > 0, \quad L_0 = l_0.
\]

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

\[
L^\delta(t) \overset{\text{def}}{=} L_0 + \int_0^t A(T(s)) ds + \int_0^t dW(s) ds,
\]

where \( T(t) = L_k \) for \( t \in [k\delta, (k + 1)\delta] \). Observe that \( L_k^\delta(k\delta) = T(k\delta) = L_k \), hence \( L^\delta(t) \) and \( T(t) \) are continuous-time extensions to the discrete-time chain \( \{L_k\}_k \). 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 \( \mathbb{E} \left[ \|L(t)\|_2^p \right] < \infty \) for any \( p \geq 2 \). Then

\[
\mathbb{E} \left[ L^\delta(T) - L(T) \right]_2 \leq \mathbb{E} \left[ \sup_{0 \leq t \leq T} \left\| L^\delta(t) - L(t) \right\|_2 \right] \to 0, \quad \delta \to 0.
\]

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**


