Agrégation à poids exponentiels : Algorithmes d'échantillonnage

Luu Duy Tung[1] Jalal Fadili[1] Christophe Chesneau[2]
[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'échantillonage 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, regularisation 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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi} \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ 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 $\boldsymbol{y}$. Generally, the problem (1) is either under-determined or determined (i.e., $p \leq n$), but $\boldsymbol{X}$ 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

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\widehat{\mu}(\boldsymbol{\theta})d\boldsymbol{\theta}, \;\; \widehat{\mu}(\boldsymbol{\theta}) \propto \exp\left(-V(\boldsymbol{\theta})/\beta\right), \tag{2}$$

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

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

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

Langevin diffusion The computation of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ 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 μ 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 $\boldsymbol{L}$ in $\mathbb{R}^p$, $p \geq 1$ is a homogeneous Markov process defined by the stochastic differential equation (SDE)

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

where $\boldsymbol{\rho} = \nabla \log \mu$, μ is everywhere non-zero and suitably smooth target density function on $\mathbb{R}^p$, $\boldsymbol{W}$ is a p-dimensional Brownian process and $\boldsymbol{l}_0 \in \mathbb{R}^p$ is the initial value. Under mild assumptions, the SDE (3) has a unique strong solution and, $\boldsymbol{L}(t)$ has a stationary distribution with density μ. This opens the door to approximating integrals $\int_{\mathbb{R}^p} f(\boldsymbol{\theta})\mu(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $f : \mathbb{R}^p \to \mathbb{R}$, by the average value of a Langevin diffusion, i.e., $\frac{1}{T}\int_0^T f(\boldsymbol{L}(t))dt$ 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

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

where $\delta > 0$ is a sufficiently small constant discretization step-size and $\{\boldsymbol{Z}_k\}_k$ are iid $\sim \mathcal{N}(0, \mathbf{I}_p)$. The average value $\frac{1}{T}\int_0^T \boldsymbol{L}(t)dt$ can then be naturally approximated via the Riemann sum $\delta/T\sum_{k=0}^{\lfloor T/\delta\rfloor - 1}\boldsymbol{L}_k$ where $\lfloor T/\delta\rfloor$ denotes the interger part of T/δ. 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 englarge the family of μ covered in [2, 5, 3, 4] by relaxing the underlying conditions. Namely, in our framework, μ is structured as $\widehat{\mu}$ with W_λ is not necessarily differentiable nor convex. Let $F_\beta = F(\boldsymbol{X}\cdot, \boldsymbol{y})/\beta$ 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

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

Define also the corresponding proximal mapping as

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

To establish the algorithm, let us state some assumptions.

(**H.1**) $W_{\beta,\lambda}$ is proper, lsc and bounded from below.

(**H.2**) $\mathrm{prox}_{\gamma W_{\beta,\lambda}}$ is single valued.

(**H.3**) $\mathrm{prox}_{\gamma W_{\beta,\lambda}}$ is locally Lipschitz continuous.

(**H.4**) $\exists K_1>0, \forall \boldsymbol{\theta}\in\mathbb{R}^p, \left\langle \boldsymbol{D}^\top\boldsymbol{\theta}, \mathrm{prox}_{\gamma W_{\beta,\lambda}}(\boldsymbol{D}^\top\boldsymbol{\theta})\right\rangle \leq K(1+\left\|\boldsymbol{\theta}\right\|_2^2)$.

(**H.5**) $\exists K_2>0, \forall \boldsymbol{\theta}\in\mathbb{R}^p, \langle \boldsymbol{\theta}, \nabla F_\beta(\boldsymbol{\theta})\rangle \leq K_2(1+\left\|\boldsymbol{\theta}\right\|_2^2)$.

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 $\mathrm{prox}_{\gamma W_{\beta,\lambda}}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

where $\overline{\boldsymbol{L}}(t) = \boldsymbol{L}_k$ for $t\in[k\delta,(k+1)\delta[$. Observe that $\boldsymbol{L}^\delta(k\delta) = \overline{\boldsymbol{L}}(k\delta) = \boldsymbol{L}_k$, hence $\boldsymbol{L}^\delta(t)$ and $\overline{\boldsymbol{L}}(t)$ are continuous-time extensions to the discrete-time chain $\{\boldsymbol{L}_k\}_k$. 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* $\mathbb{E}\left[\left\|\boldsymbol{L}(0)\right\|_2^p\right]<\infty$ *for any* $p\geq 2$. *Then*

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

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.

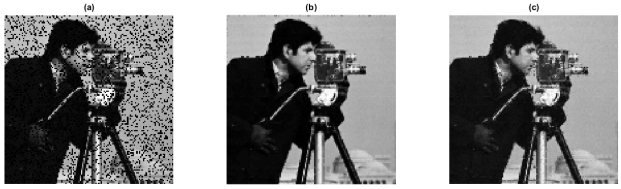

FIGURE 1 – (a) : Masked image (b) : Inpainting with EWA - $\ell_{1,2}$. (c) Inpainting with EWA - SCAD.

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.