Stochastic Approximation with Biased MCMC for Expectation Maximization

Samuel Gruffaz[1,2], Kyurae Kim[3], Jacob R. Gardner[3] and Alain Durmus[4]

[1]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190, Gif-sur-Yvette, France
[2]Université de Paris, CNRS, Centre Borelli, F-75005 Paris, France [3]University of Pennsylvania, Philadelphia, United States [4]CMAP, CNRS, École Polytechnique

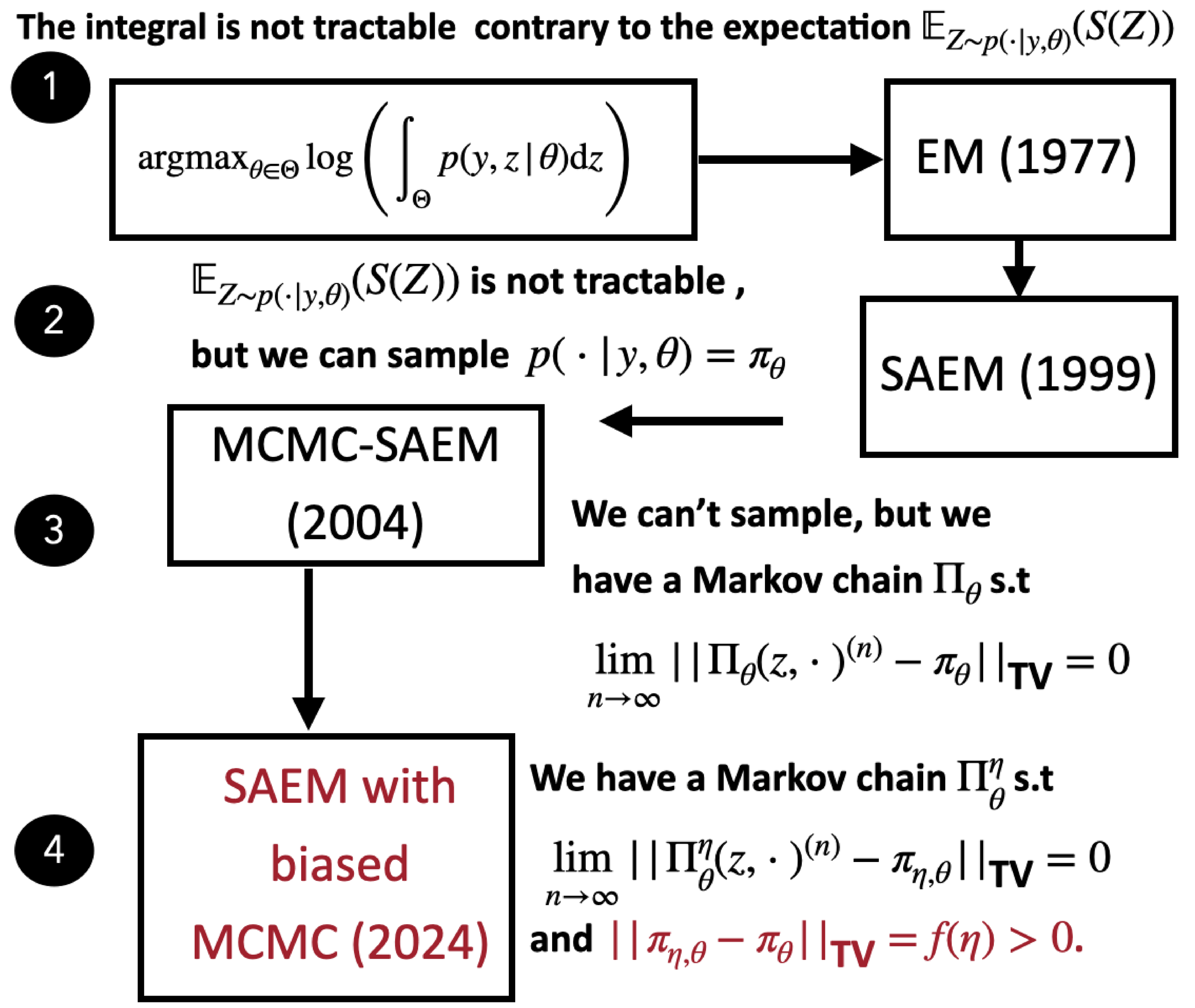
Background and Motivations

Marginal Likelihood Maximization

For a model with latent variables the empirical Bayes paradigm infers parameters by solving the maximum **marginal likelihood** problem :

$$\mathrm{argmax}_{\theta\in\Theta}\, l(\theta), \quad l(\theta) \triangleq \log\left(\int_{\mathcal{Z}} p(y,z|\theta)\mathrm{d}z\right)$$

where $z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ are latent variables, $y \in \mathcal{Y} \subset \mathbb{R}^{d_y}$ are observations, and $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ are model parameters.



ULA

The ULA Markov chain $(X_k)_{k\geq 0}$ is derived from the EulerMaruyama discretization scheme associated with the Langevin diffusion related to the force $U \triangleq -\nabla \log(\pi)$ if π is the target distribution, at iteration $k \geq 0$

$$X_{k+1} = X_k - \eta_{k+1}\nabla U(X_k) + \sqrt{2\eta_{k+1}}Z_{k+1}$$

MALA

Metropolis-Hastings-adjusted Langevin (MALA) is the asymptotically unbiased counterpart of ULA by applying a Metropolis-Hasting accept-reject step.

Contribution

We show that, for **maximum marginal likelihood** inference of **high dimensional** latent variable models, using asymptotically **biased MCMC** methods in **SAEM** is **more effective**.

Why ? According to [Durmus and Moulines(2017)], in **finite time** and **high dimension** :

$$\text{Sampling bias(ULA)} \ll \text{Sampling bias(MALA)}$$

Methodology

Jensen trick and the ELBO

Denoting by $D_{\mathcal{Z}} \triangleq \{f \in \mathrm{L}^1(\mathcal{Z}) : f \geq 0, \int_{\mathcal{Z}} f\,\mathrm{d}z = 1\}$, let $q \in D_{\mathcal{Z}}$, for any $\theta \in \Theta$,

$$-\log p(y|\theta) \leq -\mathbb{E}_{Z\sim q}(\log(p(y,Z|\theta))) + \mathbb{E}_{Z\sim q}(\log q(Z)) = -\mathrm{ELBO}(\theta, q).$$

The function $q \in D_{\mathcal{Z}} \mapsto \mathrm{ELBO}(\theta,q)$ is minimized by $q^*(z) \triangleq p(z|y,\theta)$ such that $\mathrm{ELBO}(\theta,q^*) = -l(\theta)$. Thus, by considering $\theta^* = \mathrm{argmin}_{\theta'\in\Theta}\,\mathrm{ELBO}(\theta',q^*)$, we have $l(\theta^*) \leq l(\theta)$. This procedure offers a recipe to construct a maximizing sequence of l.

H1

For any $y \in \mathcal{Y}, z \in \mathcal{Z}$ and $\theta \in \Theta$,

$$p(y,z|\theta) = h(y,z)\exp(S(y,z)^\top\phi(\theta) - \psi(\theta)),$$

Denoting by $L(s,\theta) \triangleq s\cdot\phi(\theta) - \psi(\theta)$, we define for any $s \in \mathbb{R}^d$, $\hat{\theta}(s) = \mathrm{argmax}_{\theta\in\Theta} L(s,\cdot)$. All functions are smooths.

Under **H1**, denoting by $\bar{s}(\theta) = \mathbb{E}_{Z\sim p(z|y,\theta)}(S(y,Z))$, we have, $\mathrm{argmin}_{\theta'\in\Theta}\,\mathrm{ELBO}(\theta',q^*) = \hat{\theta}\circ\bar{s}(\theta)$.

EM algorithm

Under **H1** , the EM algorithm is defined as follows : Let $(s_k)_{k\geq 0}, (\theta_k)_{k\geq 0}$ be initialized from $\theta_0 \in \Theta$ and follow the recursion for any $k \geq 0$:

1. **Expectation :** Set $s_k = \bar{s}(\theta_k)$.
2. **Maximization :** Set $\theta_{k+1} = \hat{\theta}(s_k)$, which implies $l(\theta_{k+1}) \geq l(\theta_k)$.

SAEM with biased MCMC

For any $s \in \mathbb{R}^d$ and $\eta \in (0,\eta_0]$, the Markov kernel Π_s^η has a single stationary distribution $\pi_{\hat{\theta}(s),\eta}$, also denoted as π_s^η, such that $\pi_{s,\eta}\Pi_s^\eta = \pi_{s,\eta}$.
Let $(\gamma_n)_n$, $(\eta_n)_n$ be two monotone nonincreasing sequences and for any $n \geq 0$, define the recursion,

$$Z_{n+1} \sim \Pi_{s_n}^{\eta_{n+1}}(Z_n,\cdot), \quad s_{n+1} = s_n + \gamma_{n+1}(S(y,Z_{n+1}) - s_n).$$

H2

Denoting the bias at step n by $\beta_n = \mathbb{E}_{Z\sim\pi_{\eta_{n+1},s_n}}(S(y,Z)) - \mathbb{E}_{Z\sim p(\cdot|\hat{\theta}(s_n),y}(S(y,Z))$, we have a.e $\limsup_n |\beta_n| = \beta < \infty$.

Asymptotic Theorem

Under **H1-2** and other technical conditions, on the event $A_Q = ((s_n)$ belongs to a compact $Q)$, there exists a.e K_Q s.t

$$\limsup_{n\to\infty} |\nabla V(s_n)| \leq K_Q\beta^{q/2}$$

on A_Q where $q = (p-d)/(p-1)$ if V is C^p.

Non Asymptotic Theorem

Under **H1-2** quite heavy assumptions often used in the litterature, denoting by $B(\beta) \propto \sqrt{\beta}/(\mathrm{cst} - \sqrt{\beta})$ the bias constant, with probability $1-\delta$ we have,

$$\min_{i=1,\dots,n} |h(s_i)|^2 \leq O(\log(n/\delta)/\sqrt{n} + B(\beta)).$$

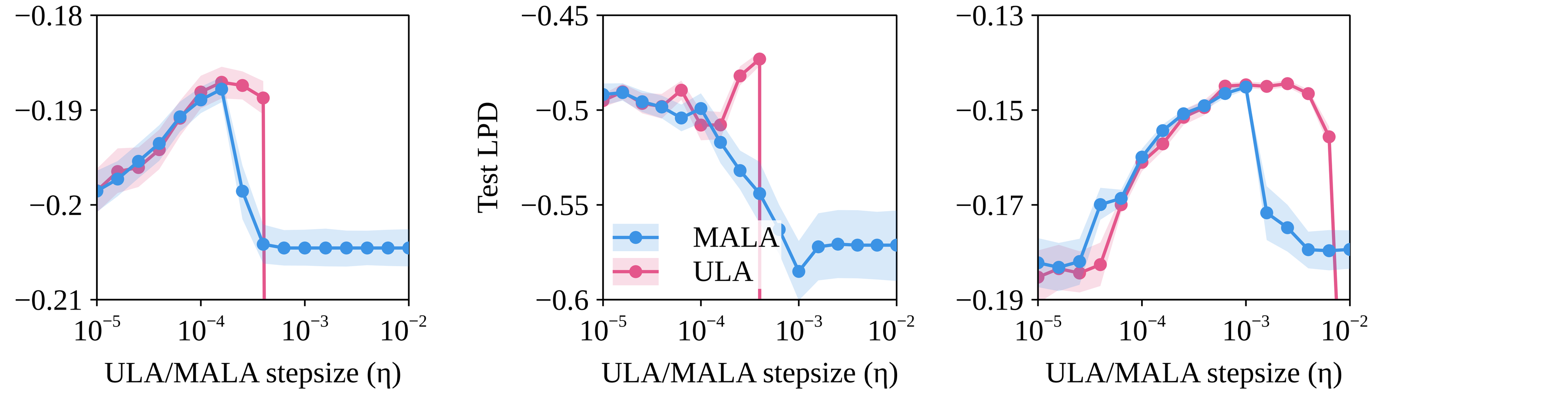
Experiments

Logistic Regression with Automatic Relevance Determination.

The model is described as :

$$\beta_0 \sim \mathcal{N}(0,10) \quad \beta \sim \mathcal{N}\left(0,\gamma^{-1}\right), \quad p_i = \mathrm{logistic}\left(\beta^\top x_i + \beta_0\right), \quad y_i \sim \mathrm{Bernoulli}\left(p_i\right),$$

where $\gamma = (\gamma_1,\gamma_2,\dots,\gamma_d) \in \mathbb{R}^d_{>0}$ is the parameter to optimize with MCMC-SAEM using MALA or ULA. We report the average log-predictive density (LPD) on 32 independant train-test split.



(a) phishing dim(z)=68 (b) german dim(z)=217 (c) caravan dim(z)=620

Bibliography

Alain Durmus and Eric Moulines.
Nonasymptotic convergence analysis for the unadjusted Langevin algorithm.
Annals of Applied Probability, 27(3) :1551–1587, 2017.

Contact : samuel.gruffaz@ens-paris-saclay.fr