Examples of Adaptive MCMC

Joel Chacón
Profesor:
Dr. José Andrés Christen García

Centro de Investigación de Matemáticas

November 25, 2019

Introduction

- Adaptive MCMC (Markov Chain Monte Carlo) can be ideal to sample complicated high-dimensional distributions (e.g statistical inference).
- Those strategies can be very successful at finding good parameter values **with little user intervention**.
- However adaptive MCMC algorithms will not always are stationary of the target distribution.

Theorem

Theorem 1.1: Ergodicity Adaptive MCMC

Consider an adaptive MCMC algorithm on a state space χ, with adaptation index Y, so $\pi(.)$ is stationary for each kernel P_γ for $\gamma \in Y$. Under the following conditions, the adaptive algorithm is ergodic.

- (Simultaneous uniform ergodicity) For all $\epsilon > 0$, there exists $N = N(\epsilon) \in \mathbb{N}$ such that $||P_\gamma^N(x, .) - \phi(.)|| \leq \epsilon$ for all $x \in \chi$ and $\gamma \in Y$.
- (Diminishing adaptation) $lim_{n \to \infty} D_n = 0$ in probability where

$$D_n = sup_{x \in \chi} ||P_{\Gamma_{n+1}}(x, .) - P_{\Gamma_n}(x, .)||$$

  is a measurable random variable (depending on the random values Γ_n and Γ_{n+1}

Conditions

Adaptive MCMC do not always preserve stationarity of the target distribution $\pi(.)$

- *Diminishing Adaptation*
  - Two successive transitions kernels are similar.

$$lim_{n \to \infty} sup_{x \in X} ||P_{\Gamma_{n+1}}(x,.) - P_{\Gamma_n}(x,.)|| = 0 \tag{1}$$

- *Bounded Convergence*
  - Ergodicity of transition kernels.

$$\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}, \quad \epsilon > 0 \tag{2}$$

Assuming those conditions the Asymptotic convergence and WLLN (Weak law of large numbers) are satisfied.

Adaptive Metropolis (AM)

Haario, Saksman, and Tamminen (2001) proposed a version of the AM algorithm.

$$Q_n(x,.) = \begin{cases} N(x,(0.1)^2 I_d/d), & \text{if } n \leq 2d \\ (1-\beta)N(x,(2.38)^2\Sigma_n/d) + \beta N(x,(0.1)^2 I_d/d), & \text{otherwise} \end{cases} \tag{3}$$

where

$\pi(,)$ is a d-dimensional target distribution

Σ_n is the current empirical estimate of the covariance of the target distribution

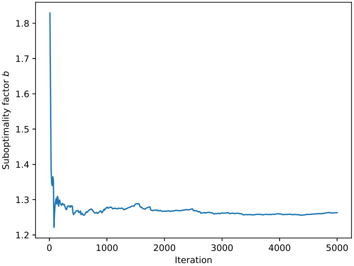
- β is a small positive constant to ensure *Boundary Convergence* ($\beta = 0.05$).
- Empirical estimates change at the nth iteration by only $O(1/n)$, *Diminishing Adaptation* is satisfied.

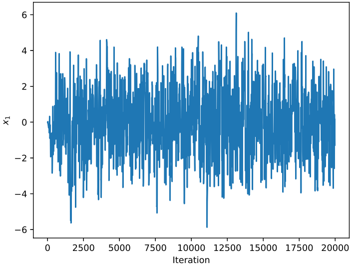
Adaptive Metropolis (AM)- Experimental validation

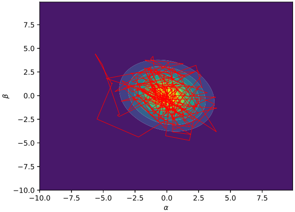
- The previously proposal was tested with $\pi(.) = N(0, MM^t)$, where $M \in \Re^{d\times d}$, $\{M_{ij}\}_{i,j=1}^{d}$ i.i.d $\sim N(0,1)$.
- This target model was configured with 2 and 100 dimensions, each one with 100 and 1'000'000 iterations correspondingly.
- The performance of the algorithm can be measured with a suboptimality factor Eqn. (4) where λ_i are the eigenvalues of $\Sigma_p^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}$.

$$b = d\frac{\sum_{i=1}^{d}\lambda_i^{-}2}{(\sum_{i=1}^{d}\lambda_i^{-1)}2} \tag{4}$$

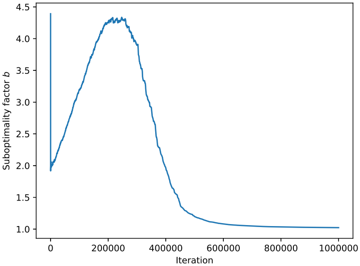
Adaptive Metropolis (AM)- Experimental validation

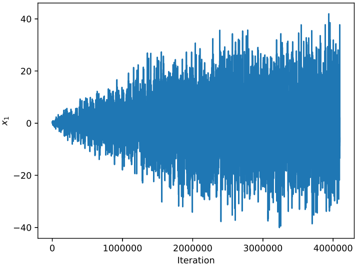






Adaptive Metropolis (AM)- Experimental validation





Adaptive Metropolis (AM)- An Irregularly Shaped Example

- AM can work well on target densities which density elliptical contours.
- Target distribution: $f_B = f_d \circ \Phi_B$, where $f_d = N(\mathbf{0}, diag(100, 1, ..., 1))$, $\Phi_B(x_1, ..., x_d) = (x_1, x_2 + Bx_1^2 - 100B, x_3, ..., x_d)^2)$.
- Taking into account 50 variables and 100’000 iterations.

$$f_B(x_1, ..., x_d) \propto exp[-x_1^2/200 - \frac{1}{2}(x_2 + Bx_1^2 - 100B)^2 - \frac{1}{2}(x_3^2 + x_4^2 + ... + x_d^2)] \quad (5)$$

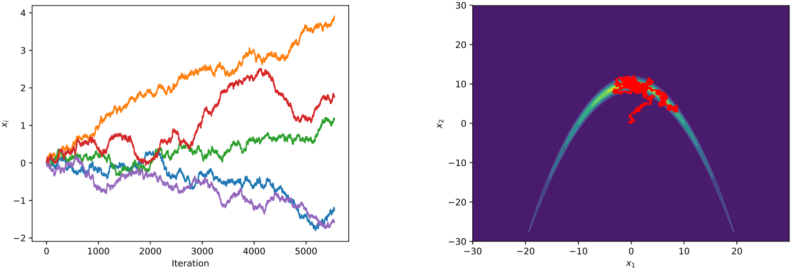
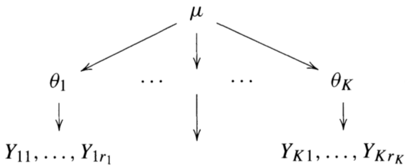


Figure 1: First five variables with 100’000 iterations and 50 variables (left side). Two dimensional contour with 1’000 iterations (right side).

Adaptive Metropolis-Within-Gibbs

Consider the following model:



where

$$
\begin{aligned}
\theta_i &\sim Cauchy(\mu, A) \quad [1 \leq i \leq K] \\
Y_{ij} &\sim N(\theta_i, V) \quad [1 \leq j \leq r_i] \\
&priors: \\
\mu &\sim N(0,1), \quad A, V \sim IG(1,1)
\end{aligned}
\tag{6}
$$

IG is the inverse gamma distribution with density propotional to $e^{-b}x^{-(a+1)}$, and Cauchy distribution is proportional to $[1 + ((x - m)/s)^2]^{-1}$.

Adaptive Metropolis-Within-Gibbs

- This model gives rise to a posterior distribution $\pi(.)$ on the $(K+3)$-dimensional vector $(A, V, \mu, \theta_1, .., \theta_k)$ conditional on the observed data $\{Y_{ij}\}$.
- We take r_i randomly from $\{5, 2, 3, 1\}$ and $K = 10$.
- The Cauchy distribution destroys conjugacy, thus classical Gibbs is infeasible.
- Test data $Y_{ij} \sim N(i-1, 10^2)$, $1 \leq i \leq K$ and $1 \leq j \leq r_i$.

Adaptive Metropolis-Within-Gibbs

- The poster distribution can be efficient-computed taking into account the logarithm.

$$f(A, V, \mu, \theta_1, .., \theta_K) \propto \left[\prod_{i=1}^{K} \prod_{j=1}^{r_i} N(Y_{ij})\theta_i V) \times Cauchy(\mu, A; \theta_1, ...m\theta_K) \right] \times N(\mu_0, \sigma_0) \times IG(A; a_1, b_1) \times IG(V; a_2, b_2) \tag{7}$$

$$log(f(A, V, \mu, \theta_1, .., \theta_K)) \propto \left[\sum_{i=1}^{K} \left(-log(1 + ((\theta_i - \mu)/A)^2) + \sum_{j=1}^{r_i} [-log(V^{0.5}) - \frac{1}{2V}(Y_{ij} - \theta_i)^2] \right) \right] - log(\sigma_0) - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 - \frac{b_1}{A} - (a_1 + 1)log(A) - \frac{b_2}{V} - (a_2 + 1)log(V) \tag{8}$$

Adaptive Metropolis-Within-Gibbs

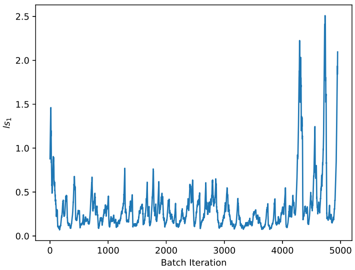
1: For each variable i $[i \leq i \leq K+3]$, create a variable ls_i giving the logarithm of the standard deviation.
2: Begin with unit variance $ls_i = 0 \; \forall i \in K$.
3: After n^{th} batch of 50 iterations update each ls_i

$$ls_i = \begin{cases} +\delta(n), & \text{if } Acceptance \quad rate \quad batch > 0.44 \\ -\delta(n), & \text{otherwise} \end{cases} \tag{9}$$

4: Choose a $\delta(n) \to 0$ to satisfy Diminishing Adaptation condition (e.g. $\delta(n) = min(0.001, n^{-0.5})$).
5: Restrict each $ls_i \in [-M, M]$ to satisfy Boundary Convergence condition.

Experimental Validation - Adaptive Metropolis-Within-Gibbs

- We take r_i randomly from $\{5, 2, 3, 1\}$ and $K = 10$.
- Batch size = 50, Batch iterations = 5000
- Test data $Y_{ij} \sim N(i-1, 10^2)$, $1 \leq i \leq K$ and $1 \leq j \leq r_i$.



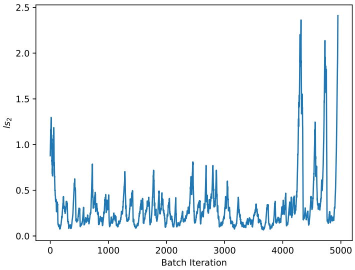


Figure 2: Log variance of the first two variables ls_1, ls_2.

Experimental Validation - Adaptive Metropolis-Within-Gibbs

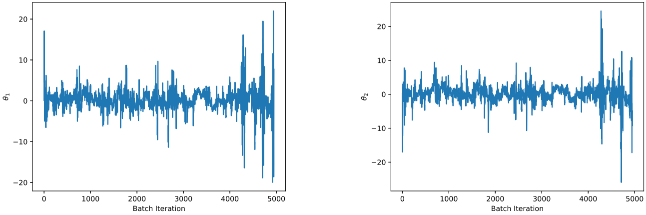


Figure 3: Values of the first two variables θ_1, θ_2.

Variable	Adaptive	Fixed
θ_1	4.27	9.98
θ_2	4.08	9.73
θ_3	4.69	10.74
θ_4	4.27	10.75
θ_5	4.31	9.63

Table 1: Avr sq. dist.

Proposal - Adaptive Weights Hybrid with Kernels

- Target distribution $\pi(.) = N(0, MM^t)$, where $M \in \Re^{d\times d}$, $\{M_{ij}\}_{i,j=1}^{d}$ i.i.d $\sim N(0,1)$.
- Proposal $\epsilon_t = w_1 N_2(0,(0.1 * I)) + w_2 N_2(0,(0.5 * I))$.
- To have more difficult $x_0 \sim U(0.0, 500.0)$.
- Each weight is implemented taking into account the acceptance ratio.
- The batch-size was set to 1000 and the number of iterations to 10'000.

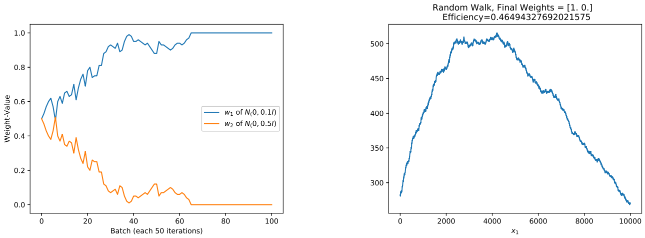


Figure 4: Weight values and states of the first variable.

Conclusions

- Adaptive MCMC provide promising results for finding good values for proposal variance, especially in cases of high dimension when it is unreasonable to do by hand.
- Adaptive strategy is too *greedy* in that it tries to adapt too closely to initial information from the output, such algorithms can take considerable time to recover from misleading initial information.
- More work should be done to design robust adaptive algorithms.
- Avoidance of *frankenstein* in the designing of adaptive algorithms is fundamental.

Acknowledgements

This work has been supported by the Center for Research in Mathematics.