Towards Scaling Up Markov Chain Monte Carlo: An Adaptive Subsampling Approach

A. Ait Oumeziane, P. Soto Martín & T. Fassina

Contents

1 Introduction

The article "Towards Scaling Up Markov Chain Monte Carlo: An Adaptive Subsampling Approach" presents a computationally efficient alternative to traditional Metropolis-Hastings sampling by leveraging subsampling. Since standard MCMC methods are computationally expensive, optimizing them is a key research area. This algorithm, designed within a Bayesian framework, approximates posterior distributions by evaluating proposed samples using only a subset of data.

The study follows a structured approach. First, it introduces the necessary mathematical notation and explains both the standard Metropolis-Hastings algorithm and its subsampling-based variant, highlighting its advantages and challenges. Next, a theoretical discussion examines a crucial aspect underlying the algorithm's validity. Finally, experiments assess its computational efficiency relative to traditional MCMC.

This analysis offers a comprehensive perspective on subsampling-based MCMC, balancing theoretical insights with practical evaluation.

Numerical experiments are available at the public github repository : `https://github.com/tizianofassina/MH_subsampling_bayesian.git`

2 MH and Subsampling MH

In this section, we present the standard Metropolis-Hastings (MH) algorithm and its subsampling-based variant in the Bayesian context. Let $p(\theta)$ be the prior density, $p(x|\theta)$ the likelihood, and $\pi(\theta)$ the posterior. The goal is to sample from $\pi(\theta)$ using a proposal kernel $q(\theta'|\theta)$.

In the classical MH algorithm, given θ_k, a new candidate θ' is proposed via $q(\theta'|\theta_k)$. The next state θ_{k+1} is set to θ' with probability

$$\alpha = \min\left(1, \frac{\pi(\theta')q(\theta_k|\theta')}{\pi(\theta_k)q(\theta'|\theta_k)}\right)$$

otherwise, the chain remains at θ_k. This criterion is equivalent to rejecting based on

$$\Lambda_n(\theta_k, \theta') > \psi(U, \theta_k, \theta')$$

where

$$\Lambda_n(\theta_k, \theta') = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p(x_i|\theta')}{p(x_i|\theta_k)}\right), \quad \psi(U, \theta_k, \theta') = \frac{1}{n}\log\left(U\frac{p(\theta_k)q(\theta'|\theta_k)}{p(\theta')q(\theta_k|\theta')}\right)$$

with $U \sim \mathcal{U}(0,1)$. Expressing α in terms of U and taking logs confirms this equivalence.

The MH algorithm iterates by initializing θ_0 and generating $(\theta_k)_{k\geq 0}$ via this rule. Under standard assumptions, this sequence converges to $\pi(\theta)$, ensuring that sufficiently many iterations yield approximate samples from $\pi(\theta)$.

A natural question arises: instead of using all data $x_1, \ldots, x_n$, can we base the acceptance decision on a subsample $x_{i_1}, \ldots, x_{i_T}$? This leads to the modified criterion

$$\Lambda_T(\theta_k, \theta') > \psi(U, \theta_k, \theta')$$

where

$$\Lambda_T(\theta_k, \theta') = \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{p(x_{i_t}|\theta')}{p(x_{i_t}|\theta_k)}\right).$$

Replacing the full-data likelihood ratio with an estimate from a subset reduces computational cost while preserving a reliable approximation of MH. We will use $\Lambda_T, \Lambda_n, \psi$ for brevity.

Before illustrating the algorithm, we introduce a useful lemma regarding real-valued inequalities. Instead of incrementing the variable t by 1 at each iteration, we increase it more aggressively to avoid unnecessary loop computations.

Lemma. Let $a, b, c, \epsilon \in \mathbb{R}$. If $|a-b| < \epsilon$, $|b-c| > \epsilon$, then

$$a > c \iff b > c.$$

This implies that if a subsample satisfies $|\Lambda_n - \Lambda_T| < c$ and $|\psi - \Lambda_T| > c$ for some c, then deciding via $\Lambda_T > \psi$ is equivalent to using the full-data criterion $\Lambda_n > \psi$, reducing computational cost. However, two challenges arise: (1) we lack Λ_n unless computed explicitly, and (2) we need an efficient subsample satisfying these inequalities.

Despite this, we can compute ψ, and Λ_T approximates Λ_n as T grows. If there exist c and δ such that $\mathbb{P}(|\Lambda_n - \Lambda_T| < c) > 1-\delta$, then if $|\psi - \Lambda_T| > c$, the decision based on $\Lambda_T > \psi$ is correct with probability at least $1-\delta$. This forms the core intuition of the algorithm.

More formally, let k index iterations, $x_1, \ldots, x_n$ be the data, u a uniform random variable, and θ' a proposal. Given a sequence $\{(c_t, \delta_t)\}_{t=1}^{n}$, $\Lambda_T(\theta_k, \theta')$ is built by uniformly sampling T points without replacement.

Suppose $\sum_t \delta_t \leq \delta$ and

$$\mathbb{P}(|\Lambda_t(\theta_k, \theta') - \Lambda_n(\theta_k, \theta')| < c_t) > 1 - \delta_t, \quad \forall t = 1, \ldots, n.$$

Then,

$$\mathbb{P}\left(\bigcap_{t=1}^{n} \{|\Lambda_t(\theta_k, \theta') - \Lambda_n(\theta_k, \theta')| < c_t\}\right) > 1 - \delta.$$

Defining

$$T = \min\left\{\inf\{t \mid |\Lambda_T(\theta_k, \theta') - \psi(u, \theta_k, \theta')| > c_t\}, n\right\},$$

we conclude that using $\Lambda_T(\theta_k, \theta') > \psi(u, \theta_k, \theta')$ ensures correctness with confidence $1 - \delta$.

A more formal treatment is omitted here as we suspect an underlying issue in this reasoning, to be explored further in theoretical discussions.

Given the previous discussion, the subsampling algorithm given a set $\{c_t, \delta_t\}$ follows this logic. Instead of incrementing the variable t by 1 at each iteration, we increase it more aggressively to avoid unnecessary loop computations.

```
Initialize $\theta_0$
for $k = 0$ to $N_{iter} - 1$ do
  Sample $\theta' \sim q(.|\theta_k)$
  $X \leftarrow \{x_1, \ldots, x_n\}$
  $X_{\text{sampled}} \leftarrow \emptyset$
  $b \leftarrow 1$
  $\Lambda \leftarrow 0$
  $t \leftarrow 0$
  $\psi \leftarrow \log\left(u \frac{p(\theta_k) q(\theta'|\theta_k)}{p(\theta') q(\theta_k|\theta')}\right)$
  Done $\leftarrow$ False
  while Done $==$ False do
    $X_{\text{batch}} \leftarrow$ sample of $b$ elements from $X \setminus X_{\text{sampled}}$
    $X_{\text{sampled}} \leftarrow X_{\text{sampled}} \cup X_{\text{batch}}$
    $\Lambda \leftarrow \frac{1}{t+b}\left(t\Lambda + \sum_{x \in X_{\text{batch}}} \log\left(\frac{p(x|\theta')}{p(x|\theta_k)}\right)\right)$
    $t \leftarrow t + b$
    Update batch size $b$ if needed
    if $|\Lambda - \psi| > c_t$ or $t \geq n$ then
      Done $\leftarrow$ True
    end if
  end while
  if $\Lambda > \psi$ then
    $\theta_{k+1} \leftarrow \theta'$
  else
    $\theta_{k+1} \leftarrow \theta_k$
  end if
end for
return $\{\theta_k\}_{k=0,\ldots,N_{iter}}$
```

3 A Potential Theoretical Issue

In the first section, we explained the intuition behind the subsampling algorithm. In particular, we made the following statement: If we can find a constant c and a value δ such that the probability of the event $|\Lambda_n - \Lambda_T| < c$ is greater than $1-\delta$, then by computing $|\psi - \Lambda_T|$, if we observe that it exceeds c, we can be confident that the correct decision is made with probability at least $1-\delta$. In other words, making the acceptance/rejection decision based on the inequality $\Lambda_T > \psi$ will lead to the correct outcome with probability $1-\delta$.

Evidently, this sentence requires a more formal explanation, which we formalize using the following general theorem.

Theorem. *(This theorem is false) Let $x, a \in \mathbb{R}$ and X be a random variable. Then, if*

$$\mathbb{P}(|X-x| > \epsilon) \geq 1-\delta,$$

we have the following implications:

$$x > a \implies \mathbb{P}(X > a \mid |X-a| \geq \epsilon) \geq 1-\delta,$$

and

$$x < a \implies \mathbb{P}(X < a \mid |X-a| \geq \epsilon) \geq 1-\delta.$$

By applying this theorem, given a step k and a subsampling step t with $X = \Lambda_t(\theta_k, \theta')$, $x = \Lambda_n(\theta_k, \theta')$, and $a = \psi(u, \theta_k, \theta')$, $\epsilon = c_t$, $\delta = \delta_t$ we obtain a formal expression of the intuition behind the algorithm (the stochasticity lies entirely in the subsampling step).

The problem is that this theorem is strictly false and so it seems to be this litteral citation from the article :

> The concentration bounds given above are helpful as they can allow us to decide whether (3) holds or not. Indeed, on the event $|\Lambda_t^*(\theta,\theta') - \Lambda_n(\theta,\theta')| \leq c_t$, we can decide whether or not $\Lambda_n(\theta,\theta') > \psi(u,\theta,\theta')$ if $|\Lambda_t^*(\theta,\theta') - \psi(u,\theta,\theta')| > c_t$ additionally holds. This is illustrated in Figure 2. Combined with the concentration inequality (6), we thus take the correct decision with probability at least $1-\delta_t$ if $|\Lambda_t^*(\theta,\theta') - \psi(u,\theta,\theta')| > c_t$.

The problem is that a proof sketch of this property/theorem would be something like (for example in the case $x > a$)

$$\mathbb{P}(X > a \mid |X-a| \geq \epsilon) \geq \mathbb{P}(X > a \cap |X-x| \leq \epsilon) \mid |X-a| \geq \epsilon)$$

$$= \mathbb{P}(X > a \mid |X-x| \leq \epsilon \cap |X-a| \geq \epsilon)\mathbb{P}(|X-x| \leq \epsilon \mid |X-a| \geq \epsilon)$$

using the Lemma we would obtain

$$= \mathbb{P}(|X-x| \leq \epsilon \mid |X-a| \geq \epsilon).$$

Using the hypothesis we would like to affirm that this probability is at least $1-\delta$, but here the probability is conditioned on the event $|X-a| \geq \epsilon$ and, of course, the implication

$$\mathbb{P}(|X-x| \leq \epsilon) \geq 1-\delta \implies \mathbb{P}(|X-x| \leq \epsilon \mid |X-a| \geq \epsilon) \geq 1-\delta$$

is false.

To illustrate our point, we present a counterexample where X is a subsample random variable and x represents the mean of a given dataset. The reasoning is straightforward. Consider three points: $x_1 = 0$, $x_2 = 1$, and $x_3 = 3$, with a fixed value $a = 0.9$. In this setup, the empirical mean is $\bar{x}_n = 1$, and if we consider a subsample of size 1, denoted as X_1, we have $X_1 \sim U_{\{x_1, x_2, x_3\}}$. Fixing $\epsilon = 0.2$, we obtain:

$$\mathbb{P}(|X_1 - \bar{x}_n| \leq \epsilon) = 0.3$$

but

$$\mathbb{P}(|X_1 - \bar{x}_n| \leq \epsilon \mid |X_1 - a| > \epsilon) = 0.$$

This suggests that the common belief that the subsampling algorithm makes correct decisions with probability at least $1 - \delta$ may not always hold. Specifically, the confidence level in the distance between Λ_n and Λ_T is valid only in the non-conditional case.

To understand when the conditional probability resembles the non-conditional one, consider the scenario where a is very close to x. In this case, the event $|X - a| \geq \epsilon$ strongly influences (or even prevents) the occurrence of the event $|X - x| \leq \epsilon$, as seen in the counterexample above. Conversely, when a is far from x, the event $|X - a| \geq \epsilon$ does not necessarily impact the probability of $|X - x| \leq \epsilon$. Consequently, the conditional probability

$$\mathbb{P}(|X - x| \leq \epsilon \mid |X - a| \geq \epsilon)$$

can be close to the non-conditional probability:

$$\mathbb{P}(|X - x| \leq \epsilon) \geq 1 - \delta.$$

To verify this intuition, we consider a dataset of $x_1, ..., x_n$ uniformly sampled from the interval $(0, 200)$. Let X_{15} be the random variable representing the empirical mean on a subsample of size 15, and let $\bar{x}_n$ denote the empirical mean of the entire dataset. The figure below illustrates the two probability surfaces:

- $\mathbb{P}(|X_{15} - \bar{x}_n| \leq \epsilon \mid |X_{15} - a| \geq \epsilon)$
- $\mathbb{P}(|X_{15} - \bar{x}_n| \leq \epsilon)$

for different values of a and ϵ. The range of a is centered around $\bar{x}_n$. We clearly observe that as a moves farther from $\bar{x}_n$, the conditional and non-conditional probabilities become more similar.

In our algorithm, the theorem should be interpreted at each iteration with $a = \psi$, $\epsilon = c_t$, $x = \Lambda_n$, and $X = \Lambda_t$. Contrary to what is stated in some articles, we believe that the acceptance rate of the algorithm is not necessarily greater than $1 - \delta$, or at least, this assertion cannot be rigorously demonstrated. The algorithm works because the values of Λ_n and ψ are sufficiently different such that the event $|\psi - \Lambda_t| \geq c_t$ does not significantly influence the probability of the event $|\Lambda_n - \Lambda_t| \leq c_t$. Consequently, we maintain a reasonable probability of making the correct decision in the subsampling algorithm, ensuring high confidence in our results.

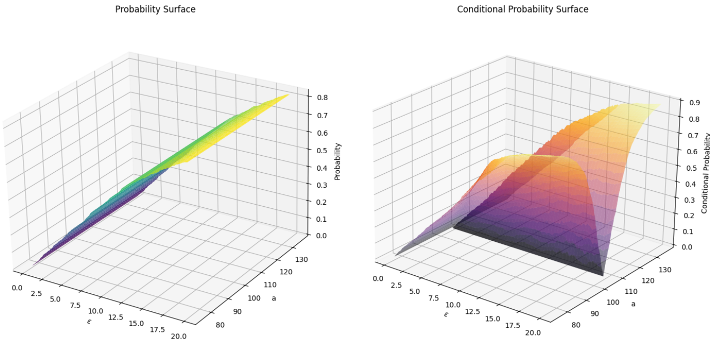
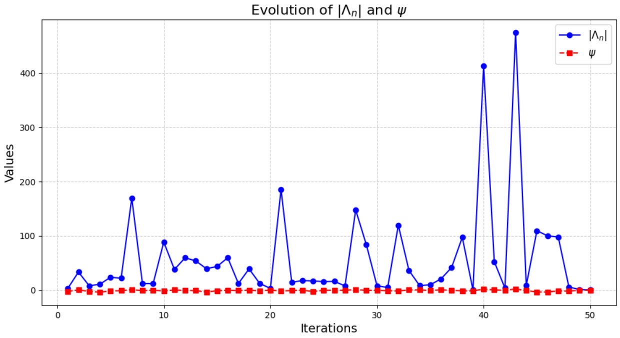


Figure 1: Comparison of conditional and non-conditional probability surfaces.

To empirically confirm this, we examine the simple case of the Γ distribution simulation using a Gaussian kernel. By observing the evolution of Λ_n and ψ across iterations (plotted in absolute value for aesthetic purposes), we confirm that their differences remain significant, thus minimizing any strong dependencies between the two probability events.



4 Numerical Experiments

We conducted a series of numerical experiments to assess the performance of the subsampling Metropolis–Hastings (MH) algorithm in realistic data scenarios. In all experiments, we compare the standard MH algorithm with its subsampling counterpart, using the hyperparameters $p = 2$, $\gamma = 2$, and $\delta = 0.01$ for the latter.

Our focus is on large-scale Bayesian logistic regression, a setting where accurately approximating the posterior distribution is often more important than minimizing classification error—particularly in applications such as Bayesian variable selection. For this reason, Bayesian logistic regression on large datasets is a natural application for the subsampling approach.

The concentration inequalities used in the algorithm rely on a constant $C_{\theta,\theta'}$, which can be explicitly computed. The log-likelihood for logistic regression is given by

$$\log p(y|x, \theta) = -\log(1 + e^{-\theta^T x}) - (1 - y)\theta^T x, \tag{1}$$

which is Lipschitz continuous in θ with Lipschitz constant $L = \|x\|$. Consequently, we define:

$$C_{\theta,\theta'} = \|\theta - \theta'\| \max_{1 \leq j \leq n} \|x_j\|. \tag{2}$$

Since equation (1) exhibits nearly linear behavior in θ, we expect this Lipschitz bound to be reasonably tight.

We employed the *covtype* dataset from [2], which contains 581,012 data points. For our experiments, we selected a training subset of n = 400,000 points, following the maximum subset size used in [2]. Although the dataset has 54 dimensions (10 quantitative), we restricted our analysis to the first $q = 2$ attributes. This choice allows us to clearly demonstrate our approach without relying on more sophisticated samplers beyond MH. We adopted the same preprocessing and Cauchy prior recommended by [3].

To assess algorithm behavior, we independently ran both standard MH and subsampling MH from four randomly chosen initializations. Figure **??** presents the evolution of the first two components of θ. In all cases, the subsampling chains converge toward values similar to those obtained by the standard MH algorithm, indicating successful posterior approximation. Moreover, the subsampling chains use, on average, only around 30% of the available data—an improvement over the results reported in [1].

Figures **??**, **??**, and **??** show the evolution of training and test errors for 2-, 15-, and 40-dimensional models, respectively. Across all dimensionalities, the test and training errors from the subsampling MH closely track those from the standard MH, even though convergence slows slightly with increased dimensionality, as expected. Importantly, the subsampling chains consistently use only about 20%–30% of the data, offering substantial computational savings. These results indicate that the subsampling MH algorithm remains effective even in higher-dimensional settings.

References

[1] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. "Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 1. Beijing, China: PMLR, 22–24 Jun 2014, pp. 405–413. URL: `https://proceedings.mlr.press/v32/bardenet14.html`.

[2] Ronan Collobert, Samy Bengio, and Yoshua Bengio. "A Parallel Mixture of SVMs for Very Large Scale Problems". In: *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker, and Z. Ghahramani. Vol. 14. MIT Press, 2001.

[3] Andrew Gelman et al. "A weakly informative default prior distribution for logistic and other regression models". In: *The Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383. DOI: `10.1214/08-AOAS191`. URL: `https://doi.org/10.1214/08-AOAS191`.