Adaptiveness of CMA based Samplers

Poster

Edna Milgo*
Vrije Universiteit Brussel
Brussels, Belgium
emilgo@vub.ac.be

Nixon Ronoh
Vrije Universiteit Brussel
Brussels, Belgium
nronoh@vub.ac.be

Peter Waiganjo
University of Nairobi
Nairobi, Kenya
waiganjo@uonbi.ac.ke

Bernard Manderick
Vrije Universiteit Brussel
Brussels, Belgium
bernard.manderick@vub.ac.be

ABSTRACT

We turn the Covariance Matrix Adaptation Evolution Strategy into an adaptive Markov Chain Monte Carlo (or *MCMC*) sampling algorithm that adapts online to the target distribution, i.e. the distribution to be sampled from. We call the resulting algorithm *CMA*-Sampling. It exhibits a higher convergence rate, a better mixing, and consequently a more effective MCMC sampler. We look at a few variants and compare their adaptiveness to a number of other adaptive samplers, including Haario et. al's *AM* sampler, on a testsuite of 4 target distributions.

CCS CONCEPTS

•Mathematics of computing → Markov-chain Monte Carlo methods; Probabilistic representations;

KEYWORDS

Adaptive Markov Chain Monte Carlo, Covariance Matrix Adaptation

ACM Reference format:
Edna Milgo, Nixon Ronoh, Peter Waiganjo, and Bernard Manderick. 2017. Adaptiveness of CMA based Samplers. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3075611

1 INTRODUCTION

Bayesian inference uses Bayes' rule to update the prior distribution of the parameters to be learned when new evidence becomes available resulting in the posterior distribution [1]. Its use, however, is hampered by two facts. One, the integrals can only be evaluated analytically for certain families of probability distributions, and two, numerical integration methods suffer from the curse of dimensionality and can only be used for low dimensional state spaces.

Monte Carlo samplers offer an alternative by generating samples according to the distribution of interest. The integral $\int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ for any function f w.r.t. to the probability measure $p(\mathbf{x})d\mathbf{x}$ can be approximated by $\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n)$ when the samples $\mathbf{x}_n$ are generated according to $p(\mathbf{x})$. However, the error in the estimate is of the order $O(1/\sqrt{N})$ where N is the number of samples. This is both good and bad news. Bad news because large amounts of samples are needed to get a small error. [1] The good news is that this estimate is independent of the dimension of the space on which the probability distributions are defined. As a result, Markov chain Monte Carlo (MCMC) samplers are the only remaining alternative when we deal with arbitrary high-dimensional probability distributions.

Most MCMC samplers can be seen as an extension of the Metropolis Hastings (MH) algorithm. Their performance critically depends on the proposal distribution used and requires a lot of tuning to find the optimal one for the problem at hand. Adaptive Metropolis samplers were introduced for this reason, with the main goal of adaptation being to improve mixing in the chain [2]. This was achieved updating online the covariance matrix of the Gaussian proposal distribution.

Covariance Matrix Adaptation Evolution Strategies [3] is the state of the art in black box stochastic optimization. It continuously adapts the covariance matrix of the search distribution to generate better offspring. In recent work we have shown how its most basic variant (1+1)-CMA can be used to adapt the covariance matrix of the proposal distribution and that the resulting adaptive sampler outperforms other ones on a standard test-suite [4].

The rest of the paper is organized as follows. In Section 2, we review the samplers that we incorporated in the comparison. In Section 3 we describe the test-suite used and the experiment done before we conclude and describe future work in Section 4.

2 MCMC SAMPLERS

The samplers compared are 1) Metropolis-Hastings with optimal proposal distribution (MH), 2) the Adaptive Metropolis (AM) algorithm, and 3) two variants of CMA based Sampling (CMA) [1–3].

*E. Milgo and N.Ronoh wish to thank VLIR-UOS, Vrije Universiteit Brussel and Moi University without whose support this work would not have been possible.

[1] For instance, to estimate π up to a precision of 10^{-10} at least 10^{20} samples are needed. This led Alan Sokal, a well-known physicist doing research on MC samplers, to the quote "Monte Carlo is is an extremely bad method; it should be used only when all alternative methods are worse."

GECCO '17 Companion, Berlin, Germany
© 2017 Copyright held by the owner/author(s). 978-1-4503-4939-0/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3067695.3075611

All samplers can be seen as extensions of the basic MH-algorithm described next.

MH generates samples from the *target* distribution $\pi(\mathbf{x})$ in two steps. First, it uses a proposal distribution (multivariate Gaussian distribution) with mean vector $\mathbf{x}_n$ and covariance matrix $\mathbf{C}_n$ to generate a candidate $\mathbf{x}^* \sim N(\mathbf{x}_n, \mathbf{C}_n)$ where $\mathbf{x}_n$ is the current sample. Next, the MH-acceptance criteria is used to decide whether the proposed $\mathbf{x}^*$ or the current $\mathbf{x}_n$ becomes the next sample $\mathbf{x}_{n+1}$. In case of a symmetric distribution like the multivariate Gaussian, it states that the candidate $\mathbf{x}^*$ is accepted with probability $\alpha(\mathbf{x}_n, \mathbf{x}^*) = \min\{1, \pi(\mathbf{x}^*)/\pi(\mathbf{x}_n)\}$. This criteria ensures $\mathbf{x}_{n+1} \sim \pi(\mathbf{x})$ whenever $\mathbf{x}_n \sim \pi(\mathbf{x})$, i.e, samples are generated according to the target distribution when the chain has converged. Many challenges impede MH from its maximum potential. We focus on *mixing*. In order to give reliable estimates, the chain has to explore evenly all areas that contribute significantly to the probability mass of the target distribution. When this is the case the chain is said to mix well. A good adaptive sampler mixes well rapidly and we study the adaptiveness of some samplers.

Basically, adaptive MCMC is MH where the Gaussian proposal distribution is adaptive, i.e. the proposed candidate $\mathbf{x}^* \sim N(\mathbf{x}_n, \mathbf{C}_n)$ where $\mathbf{C}_n$ changes over time. Adaptive samplers differ in the global scales and covariances used and how they are updated.

In case of AM, the global scale remains fixed to $\sigma = 1$ and the covariance matrix is updated as $\mathbf{C}_{n+1} = (1 - \frac{1}{n+1})\mathbf{C}_n + \frac{1}{n+1}(\mathbf{x}_{n+1} - \mathbf{m}_n)(\mathbf{x}_{n+1} - \mathbf{m}_n)^\top$ where $\mathbf{m}_n \triangleq \sum_{i=1}^{n} \mathbf{x_n}$ is the sample mean of the previous states.

Algorithm 1 shows the generic algorithm CMA Sampler. We considered $(1+1)$-CMA and $(\mu, \lambda)-$CMA. In all cases we use the recommended parameter settings as given in [3, 5, 6]. CMA uses more complex update rules than AM. But these rules guarantee that CMA-ES is invariant under general linear transformations, cf. Proposition 9 in [3] that also motivates the update rules and gives the default parameter settings.

Algorithm 1: CMA Sampling

```
repeat
    generate parents $\mathbf{z}_i \sim N(\mathbf{x}_n, \mathbf{C}_n)$ for $i = 1, \cdots, \lambda$
    recombine the $\mu$ best ones to get $\mathbf{x}^*$
    $u \sim U([0, 1])$
    if $u <= \alpha(\mathbf{x}_n, \mathbf{x}^*, \pi)$, cf. Eq. 2 then $\mathbf{x}_{n+1} \leftarrow \mathbf{x}^*$
    else $\mathbf{x}_{n+1} \leftarrow \mathbf{x}_n$
    update the global scale $\sigma_n$
    update the covariance $C_n$
until stopping criterion is met
```

3 EXPERIMENTS

We use the suboptimality factor b as performance criterion to compare the adaptiveness of the samplers considered:

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

where d is the dimension of the state space and the λ_i are the eigenvalues of the matrix $C_n^{1/2} C_{target}^{-1/2}$. C_n is the covariance matrix of the proposal distribution at step n. Note that b is constant for the MH-sampler. The closer $b > 1$ is to 1, the better the mixing of the chain, cf. [7]

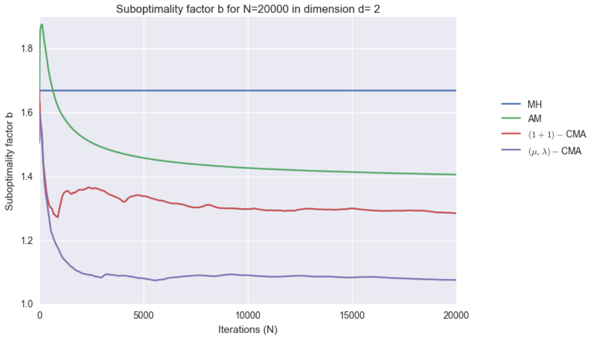

Figure 1: Suboptimality factor b for the four samplers for dimension d =100 π_3 target distribution. The CMA algorithms adapt faster.

In Figure 1 we have plotted b for all samplers for target π_3 in the testsuite used in [2]. It consists of 4 target distributions defined over $\mathbb{R}^d$ that are increasingly more challenging for samplers. The results shown are averaged over 100 independent runs. Here, we only report on one of the twisted distributions in dimension $d = 100$ since these are the most difficult ones to sample from and the potential benefit of adaptive MCMC is the highest. The results for the other targets and dimensions d are similar.

4 CONCLUSION

Although we used the strategic parameter settings originally proposed for optimization, adaptive CMA samplers show promising results. Our working hypothesis is that the invariance properties of CMA explain its better adaptiveness [3]. Future work will open perspectives for population MCMC samplers.

REFERENCES

[1] Steve Brooks, Andrew Gelman, Galin L. Jones, and XiaoLi Meng. *Handbook of Markov Chain Monte Carlo (Chapman & Hall/CRC Handbooks of Modern Statistical Methods)*. Chapman & Hall CRC, 2011.

[2] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7 no. 2:223–242, 2001.

[3] Anne Auger and Nikolaus Hansen. Tutorial cma-es: Evolution strategies and covariance matrix adaptation. In *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '12, pages 827–848, New York, NY, USA, 2012. ACM.

[4] E. Milgo, N. Ronoh, P. Waiganjo, and B. Manderick. Comparison of adaptive mcmc samplers. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2017.

[5] Graheme A. Jastrebski and Dirk V. Arnold. Improving Evolution Strategies through Active Covariance Matrix Adaptation. In *IEEE Congress on Evolutionary Computation – CEC 2006*, pages 2814–2821, 2006.

[6] Dirk V. Arnold and Nikolaus Hansen. Active covariance matrix adaptation for the (1+1)-cma-es. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, GECCO '10, pages 385–392, New York, NY, USA, 2010. ACM.

[7] Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, Vol.18(No.2), 2009.