Natural gradient evolution strategies for adaptive sampling

Conference Paper · July 2022

DOI: 10.1145/3520304.3533939

Nixon Ronoh
Vrije Universiteit Brussel
7 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Edna Milgo
Moi University
5 PUBLICATIONS 3 CITATIONS
SEE PROFILE

Ambrose Kiprop
Moi University
64 PUBLICATIONS 641 CITATIONS
SEE PROFILE

Natural Gradient Evolution Strategies for Adaptive Sampling

Nixon Ronoh*
Edna Milgo*
nixon.ronoh@vub.be
emilgo@vub.ac.be
Vrije Universiteit Brussel
Brussels, Belgium

Ambrose Kiprop
Moi University
Eldoret, Kenya
ambkiprop@gmail.com

Bernard Manderick
Ann Nowe
bmanderi@vub.be
ann.nowe@vub.be
Vrije Universiteit Brussel
Brussels, Belgium

ABSTRACT

We evaluate two (1+1)-natural evolution strategies (NES) turned into adaptive Markov chain Monte Carlo (MCMC) samplers on a test suite of probability distributions. We compare their performance with the AM-family of samplers considered to be the state of the art in adaptive MCMC. Our experiments show that natural gradient based adaptation used in NES further improves adaptive MCMC.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Markov decision processes**;

KEYWORDS

Adaptive MCMC, evolution strategies, natural gradient

ACM Reference Format:
Nixon Ronoh, Edna Milgo, Ambrose Kiprop, Bernard Manderick, and Ann Nowe . 2022. Natural Gradient Evolution Strategies for Adaptive Sampling. In *Genetic and Evolutionary Computation Conference Companion (GECCO '22 Companion), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3520304.3533939

1 INTRODUCTION

Bayesian inference generates the posterior from the prior and the evidence. Often this requires the evaluation of integrals defined over high dimensional spaces. This calls for efficient estimation methods, a gap readily filled by the Metropolis-Hastings (MH) and other Markov chain Monte Carlo (MCMC) methods that generate samples from the posterior. MH starts with a random initial sample and proceed as follows. A simpler *proposal* distribution centred in the current sample is used to generate a candidate sample. Next, we decide whether to move to that candidate or stay where we currently are. This decision is such that when the current sample is from the posterior this is also the case for the candidate. This procedure continues till the chain converges. Once converged, the chain generates samples from the posterior. We collect enough of them to estimate reliable statistics of the posterior.

The standard error of the estimate depends on 1) how well the chain mixes and 2) the correlation among generated samples. Depending on the proposal distribution the chain can either explore the state space too slowly or stagnate for longer periods. Adaptive MCMC attempts to address these problems by automatically tuning the parameters of the proposal distribution while sampling.

The covariance matrix adaptation evolution strategy (CMA ES) is state of the art stochastic optimization [6]. It adapts the parameters, i.e. mean, scale and covariance, of a Gaussian search distribution to optimize the search process. Natural Evolution Strategies (NES) [8] improve on CMA ES. It defines a differential manifold on the space of feasible parameters of the Gaussian search distribution with the Fisher information as Riemannian metric. NES uses the gradient w.r.t. to that metric, called the natural gradient, to update the parameters.

We investigate the performance of NES-inspired adaptive MCMC on a test suite of distributions. The rest of this paper is organized as follows. Section 2 describes the experimental set up, discusses the benchmarks in the test suite and describes the experiment context. Section 3 provides a summary of the experimental results. We finally conclude in Section 4.

2 EXPERIMENTAL SET UP

Two variants of NES-inspired samplers are considered: (1+1)-exponential and (1+1)-separable NES abbreviated as (1+1)-xNES and (1+1)-sNES, respectively. In both variants, referred to as (1+1)-x/sNES, one parent generates one offspring in each generation. Both compete to survive in the next generation.

xNES uses an exponential mapping to parametrize the covariance of the search distribution [8]. This makes it possible to update the scale and covariance of a distribution[1] in the *natural coordinate system* of the Riemannian manifold. Algorithm 1 gives the pseudocode. The function $J(\sigma, L) \triangleq \mathbb{E}_Q\left[\pi(\mathbf{x})\right] = \int_{\mathbb{R}^d} \pi(\mathbf{x}) Q\left(\mathbf{x}; \sigma, L\right)$ is the expectation of the target $\pi(\mathbf{x})$ w.r.t. the current proposal $Q\left(\mathbf{x}; \sigma, L\right)$ where $\Sigma = \sigma L L^\top$ is the covariance of Q.

sNES is an alternative to xNES that constrains the covariance to be diagonal [8]. This reduces computational complexity during updates.

The performances of (1+1)-x/sNES are compared with that of the well known Adaptive Metropolis (AM) algorithm introduced in [5].

Preliminary experiments consist of 4 sets of 100 runs per sampler, each set being for the state space dimensions $d = 4, 10, 50$ and 100.

The test suite consists of Haario's benchmarks π_1, π_2, π_3, and π_4 [5], and Neal's funnel distribution π_{Neal} [7].

[*] First two authors contributed equally to this research.

[1] In case of sampling, the mean is not updated.

GECCO '22 Companion, July 9–13, 2022, Boston, MA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9268-6/22/07.
https://doi.org/10.1145/3520304.3533939

Algorithm 1: (1+1)-xNES sampler

```
Input  : Initial sample $\mathbf{x}_0$, scale $\sigma_0$ and $L_0$ s.t. $\Sigma_0 = \sigma_0 L_0 L_0^\top$
         is the covariance of $Q$
Output : Samples $(\mathbf{x}_n)_{n=0}^{N}$
repeat
    Generate candidate $\mathbf{x}^* = \mathbf{x}_n + \sigma_n L_n \mathbf{z}$ where $\sigma_n$ and $L_n$
      are the current scale and Cholesky factor, and
      $\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$
    Decide, using the Metropolis criterion, whether $\mathbf{x}^*$ or $\mathbf{x}_n$
      becomes $\mathbf{x}_{n+1}$
    Compute the natural gradients $\nabla_\sigma J$, $\nabla_L J$
    Update $\sigma_{n+1} = \sigma_n \exp(\eta_\sigma \nabla_\sigma J)$, $L_{n+1} = L_n \exp(\eta_L \nabla_L J)$
until stopping criterion is met
```

Benchmark π_1 is the d-dimensional uncorrelated Gaussian distribution with covariance $\Sigma_1 = diag(100, 1, 1 \ldots 1)$ while the correlated covariance Σ_2 of the π_2 is a rotation of Σ_1. Both π_3 and π_4 twist π_1 according to parameter $b > 0$ in the transformation $\Phi_b(x_1, x_2, \cdots, x_d) = (x_1, x_2 + b(\mathbf{x}_1 - 100), x_3, \ldots, x_d)$. This twist is moderate for $b = 0.03$ while high for $b = 0.1$.

Benchmark π_{Neal} is a funnel shaped distribution. The funnel is very narrow but has very high probability mass making it difficult to generate samples in that region. According to [7] π_{Neal} is typical for posteriors in hierarchical Bayesian models and is known to be one of the most difficult to sample from. Non-adaptive samplers face problems on these benchmarks. The test suite is ideal to evaluate and/or compare of adaptive samplers.

Each run of experiments involves the *set up* of the *experiment context*: the *dimension d* of the *state space*, the *sampler* under consideration, and the number N of samples to be generated. The context is used to generate the corresponding d-dimensional state space and test suite, and to determine the optimal control parameters of that sampler, e.g. the optimal scale $\sigma_{opt} = 2.38/\sqrt{d}$ as described in Gelman et al. [3]. Subsequently, the sampler with the control parameters as arguments is run on the distributions in the test suite. The generated samples are compared with i.i.d. (independent and identically distributed) samples from the same distribution. The performance is analyzed using scatter and trace plots, cf. Figures 2 and 1. Also, the effective sample size N_{eff} [2], the $\hat{R}$-convergence measure, and the overhead running times are determined [1, 2, 4].

3 SUMMARY OF EXPERIMENTAL RESULTS

The objectives of the experiments are twofold. One, to compare the performance given specific target distributions in their order of complexity. Two, to compare the performance of (1+1)-x/sNES relative to the well known AM algorithm. Various methods were jointly used to either detect the lack of, or give an indication of convergence of the MCMC chain, namely Gelman Rubin's $\hat{R}$, effective sample size, auto-correlation and multi-chain trace plots.

The scatter plot in Figure 1 compares the first 2 coordinates of samples generated by (1+1)-x/sNES with i.i.d samples. Figure 2 shows 1-d traces of the first coordinate of samples generated ($N = 50,000$) by AM, xNES and sNES, respectively.

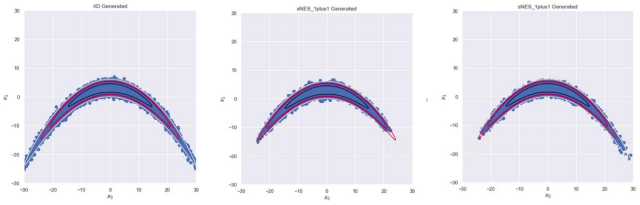

Figure 1: Scatter plot of first 2 coordinates of samples of target π_3 that are i.i.d. (left), generated by xNES (middle), and sNES (right).

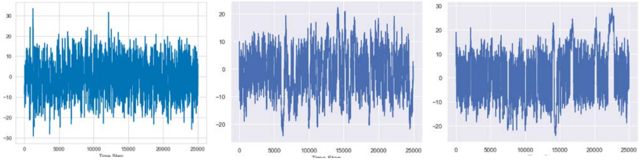

Figure 2: Trace plot of the first coordinate of samples of π_3 generated by AM (left), xNES (middle), sNES (right).

The experimental results are explained by the fact that NES exploit the geometry of the parameter space to improve adaptiveness. Based on these results and the discussion above, we now conclude.

4 CONCLUSION

The experimental results have firstly shown that the proposed (1+1)-x/sNES are indeed valid samplers, and with a reward of relatively good performance. Their sample distribution is relatively similar to that of i.i.d. samples. The competitive performance of (1+1)-x/sNES is also observed for more complex distributions.

Secondly, our investigations reveal that (1+1)-x/sNES complement the Adaptive Metropolis (AM) algorithm. This viewpoint is supported by the fact that they consistently sample from high probability areas of the distribution. This not always the case for AM. Further, (1+1)-x/sNES post higher acceptance rates than AM. It is also noted that even for very high dimensions, the high probability areas are discovered and explored while the performance of AM significantly weakens. Finally, the convergence diagnostic tests also returned competitive results.

REFERENCES

[1] Stephen P. Brooks and Andrew Gelman. 1998. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 7-4 (1998), 434–455.

[2] Samantha R Cook, Andrew Gelman, and Donald B Rubin. 2006. Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics* 15-3 (2006), 675–692.

[3] Andrew Gelman, Gareth O. Roberts, and Walter R. Gilks. 1996. Efficient Metropolis jumping rules. In *Bayesian Statistics*, J. M. Bernado et al. (Eds.). Vol. 5. 599–607.

[4] Andrew Gelman and Donald Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statist. Sci.* 7 (1992), 457–511.

[5] Heikki Haario, Eero Saksman, and Johanna Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli* 7-2 (2001), 223–242.

[6] Nikolaus Hansen. 2010. The CMA Evolution Strategy: A Tutorial. (2010). http://www.lri.fr/~hansen/cmatutorial.pdf

[7] Radford Neal. 2003. Slice sampling. *Annals of Statistics* 31-3 (2003), 705–767.

[8] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. 2014. Natural Evolution Strategies. *Journal of Machine Learning Research* 15 (2014), 949–980. http://jmlr.org/papers/v15/wierstra14a.html

[2] N MCMC samples contain the same information as N_{eff} i.i.d. samples, usually $N_{eff} << N$.