AdvancedHMC.jl:
A robust, modular and efficient implementation of advanced HMC algorithms

Kai Xu  
University of Edinburgh  
KAI.XU@ED.AC.UK

Hong Ge  
University of Cambridge  
HG344@CAM.AC.UK

Will Tebbutt  
University of Cambridge  
WCT23@CAM.AC.UK

Mohamed Tarek  
UNSW Canberra  
M.MOHAMED@STUDENT.ADF.A.EDU.AU

Martin Trapp  
Graz University of Technology  
MARTIN.TRAPP@TUGRAZ.AT

Zoubin Ghahramani  
University of Cambridge & Uber AI Labs  
ZOUBIN@CAM.AC.UK

Abstract
Stan’s Hamilton Monte Carlo (HMC) has demonstrated remarkable sampling robustness and efficiency in a wide range of Bayesian inference problems through carefully crafted adaption schemes to the celebrated No-U-Turn sampler (NUTS) algorithm. It is challenging to implement these adaption schemes robustly in practice, hindering wider adoption amongst practitioners who are not directly working with the Stan modelling language. AdvancedHMC.jl (AHMC) contributes a modular, well-tested, standalone implementation of NUTS that recovers and extends Stan’s NUTS algorithm. AHMC is written in Julia, a modern high-level language for scientific computing, benefiting from optional hardware acceleration and interoperability with a wealth of existing software written in both Julia and other languages, such as Python. Efficacy is demonstrated empirically by comparison with Stan through a third-party Markov chain Monte Carlo benchmarking suite.

1. Introduction
Hamiltonian Monte Carlo (HMC) is an efficient Markov chain Monte Carlo (MCMC) algorithm which avoids random walks by simulating Hamiltonian dynamics to make proposals (Duane et al., 1987; Neal et al., 2011). Due to the statistical efficiency of HMC, it has been widely applied to fields including physics (Duane et al., 1987), differential equations (Kramer et al., 2014), social science (Jackman, 2009) and Bayesian inference (e.g. Bayesian neural networks; Neal, 2012). The No-U-Turn Sampler (NUTS; Hoffman and Gelman, 2014) is an extension of the HMC sampler which automatically tunes two key parameters, the leapfrog step size and integration time (aka trajectory length), which used to require manual adjustments through extensive preliminary runs. Together with a robust implementation in
the Stan probabilistic programming language (PPL), NUTS has become the default choice for HMC sampling for many probabilistic modelling practitioners (Carpenter et al., 2017).

Although the integration of NUTS in Stan makes Bayesian inference easy for domain experts relying on the Stan language, it is desirable to have a high quality, standalone NUTS implementation in a high-level language, e.g. for research on HMC algorithms, reproducible comparisons and real-world approximate inference applications. To this end, we introduce AdvancedHMC.jl (AHMC), a robust, modular and efficient implementation of Stan’s NUTS and several other commonly used HMC variants in Julia.\textsuperscript{1}

2. A modular Hamiltonian Monte Carlo implementation

AHMC supports various HMC algorithms in the set below resulted from a Cartesian product of a set of HMC trajectories and a set of adaptors:

\[(\text{StaticTrajectory} \cup \text{DynamicTrajectory}) \times \text{Adaptor}\,.
\]

Here StaticTrajectory refers to a set of HMC with fixed-length trajectory length, which contains HMC with fixed step size and step numbers and HMC with fixed total trajectory length. DynamicTrajectory is a set of HMC with adaptive trajectory length which is defined as a Cartesian product of four sets of different HMC components:

Metric × Integrator × TrajectorySampler × TerminationCriterion,

where

\[
\begin{align*}
\text{Metric} &= \{\text{UnitEuclidean, DiagEuclidean, DenseEuclidean}\} \\
\text{Integrator} &= \{\text{Leapfrog, JitteredLeapfrog, TemperedLeapfrog}\} \\
\text{TrajectorySampler} &= \{\text{Slice, Multinomial}\} \\
\text{TerminationCriterion} &= \{\text{ClassicNoUTurn, GeneralisedNoUTurn}\}
\end{align*}
\]

Finally, Adaptor consists of any BaseAdaptor or any composition of two or more BaseAdaptor, where BaseAdaptor ∈ \{Preconditioner, NesterovDualAveraging\}. A special composition called StanHMCAdaptor is provided to compose Stan’s windowed adaptor, which has been shown to be robust in practice (Carpenter et al., 2017).

2.1. Example code of building Stan’s NUTS using AdvancedHMC.jl

The code snippet below illustrates the use of AHMC for a given target log density function and its gradient, logdensity\_f and grad\_f respectively.

\[1\text{ using AdvancedHMC}\]
\[2\text{n samples, n adapts, target} = 10,000, 2,000, 0.8 \quad \# \text{set up sampling parameters}\]
\[3\text{q0} = \text{randn(D)} \quad \# \text{draw a random starting point}\]
\[4\text{### Building up NUTS}\]
\[5\text{metric} = \text{DiagEuclideanMetric}(D) \quad \# \text{diagonal Euclidean metric space}\]
\[6\text{h} = \text{Hamiltonian(metric, logdensity\_f, grad\_f)} \quad \# \text{Hamiltonian on the target distribution}\]
\[7\text{eps\_init} = \text{find\_good\_eps}(h, q0) \quad \# \text{initial step size}\]

\textsuperscript{1} Code is available at https://github.com/TuringLang/AdvancedHMC.jl.
```python
int = Leapfrog(eps_init)  # leapfrog integrator
traj = NUTS(Multinomial, GeneralisedNoUTurn)(int)  # multi. sampling with gen. no U-turn
adaptor = StanHMCAdaptor(  # Stan's windowed adaptor
    n_adapts, Preconditioner(metric), NesterovDualAveraging(target, eps_init)
)
samples, stats = sample(h, traj, q0, n_samples, adaptor, n_adapts)  # draw samples
```

Here \( \logdensity_f \) and \( \text{grad}_f \) are functions of the target distributions’s log density and its gradient, which are functions independent of \text{AHMC} and can be, e.g. derived from different PPLs or defined by normalising flows (Rezende and Mohamed, 2015; Dinh et al., 2016; Papamakarios et al., 2017). We will show an example of such models in the next section.

### 2.2. GPU support for AdvancedHMC.jl via CuArrays.jl

In order to run HMC on CUDA, one only needs to change Line 3 of the demo code from \( q_0 = \text{randn}(D) \) to \( q_0 = \text{CuArray}(\text{randn}(D)) \), assuming \( \logdensity_f \) and \( \text{grad}_f \) in Line 6 are GPU friendly, which is how CuArrays.jl could be used with AdvancedHMC.jl to run HMC on GPUs. An example using GPU accelerated \text{AHMC} to draw samples from a normalising flow, named \text{flow model}, trained on MNIST (LeCun, 1998) is given below. The function \( \logpdf(m, x) \) is used to compute the log density for model \( m \) on data batch \( x \).

```python
1 \logdensity_f(x) = \logpdf(\text{flow model}, \text{reshape}(x, 784, 1))[1]
2 # Define gradient function via reverse AD
3 function grad_f(x)
4    val, back = Tracker.forward(logdensity_f, x)
5    grad = back(1)
6    return (Tracker.data(val), Float32.(Tracker.data(grad[1][:,1])))
7 end
```

Here Tracker is an automatic differentiation (AD) library which implements reverse-mode AD. *How does it work?* All arrays in \text{AHMC} are abstractly typed, meaning that the concrete type is deduced at compile time from \( q_0 \). That is to say, if \( q_0 \) is defined on the GPU, i.e. it is a \text{CuArray}, all internal arrays in HMC will be too.

### 3. Related work

A summary of the related work on HMC/NUTS implementations is given in Table 3. We want to emphasis that there exists a Python API of \text{AHMC} implemented by an independent team available at https://github.com/salilab/hmc.

### 4. Evaluations

To compare the NUTS implementation between \text{AHMC} and \text{Stan}, we consider five models from MCMCBenchmarks.jl, a package for benchmarking MCMC libraries in Julia.\(^2\) The five models are 1) a *Gaussian model* (Gaussian), a simple two parameter Gaussian distribution, 2) the *signal detection model* (SDT), a model used in psychophysics and signal processing (Green et al., 1966), 3) a *linear regression model* (LR), 4) a *hierarchical Poisson regression* (HPR) model and 5) the *linear ballistic accumulator model* (LBA), a cognitive model.

\(^2\) Available at https://github.com/StatisticalRethinkingJulia/MCMCBenchmarks.jl.
Table 1: Comparison of different HMC/NUTS implementations. **TFP.MCMC** refers to **TensorFlow.Probability**’s MCMC library. **DynamicHMC** is another high-quality HMC implementation in Julia. Windowed adaption is a method for joint adaption of leapfrog integration step-size and mass matrix. Windowed adaption was proposed by the Stan team (Carpenter et al., 2017), and has demonstrated remarkable robustness in a wide range of Bayesian inference problems. Partial support for GPU means the log density function can be accelerated by GPU, but the HMC sampler itself runs on CPU. Slice and Multinomial are two methods for sampling from dynamic Hamiltonian trajectories, e.g. those generated by the No-U-Turn algorithm (see e.g. Betancourt (2017) for details). Tempered leapfrog integrator improves convergence when the target distribution has multiple modes by performing tempering within Hamiltonian trajectories (Neal et al., 2011).
Figure 1: LR (50 runs). For the density plots, blue is for AHMC and orange is for Stan and each row is for a different size \( N \), corresponding to the table on the right.

<table>
<thead>
<tr>
<th>( N )</th>
<th>Gaussian ( ^2 )</th>
<th>SDT ( ^3 )</th>
<th>LR ( ^2 )</th>
<th>HPR ( ^1 )</th>
<th>LBA ( ^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_0 )</td>
<td>( \sigma )</td>
<td>( b_1 )</td>
<td>( b_2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stan</td>
<td>10</td>
<td>0.8039</td>
<td>10</td>
<td>0.7759</td>
<td>10</td>
</tr>
<tr>
<td>AHMC</td>
<td>10</td>
<td>0.3361</td>
<td>10</td>
<td>0.3285</td>
<td>10</td>
</tr>
<tr>
<td>Stan</td>
<td>100</td>
<td>0.7561</td>
<td>100</td>
<td>0.7261</td>
<td>100</td>
</tr>
<tr>
<td>AHMC</td>
<td>100</td>
<td>0.3303</td>
<td>100</td>
<td>0.3201</td>
<td>100</td>
</tr>
<tr>
<td>Stan</td>
<td>1000</td>
<td>0.7614</td>
<td>1000</td>
<td>0.7089</td>
<td>1000</td>
</tr>
<tr>
<td>AHMC</td>
<td>1000</td>
<td>0.5081</td>
<td>1000</td>
<td>0.3179</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 2: Computational efficiency for five models using \(^1\) 25 runs, \(^2\) 50 runs or \(^3\) 100 runs. For AHMC, forward-mode AD is used for computing gradient. AHMC can be used together with different AD backends/packages, e.g. ForwardDiff.jl’s forward-mode AD, Tracker.jl’s reverse-mode AD and Zygote.jl’s source-to-source AD.

4.2. Computational efficiency via running time

All the benchmark models used in this paper are implemented in Turing (Ge et al., 2018), a universal PPL in Julia that uses AHMC as its HMC backend; see Appendix C for an example Turing code. The average time used to run the five benchmark models for multiple times using Stan and using AHMC via Turing are reported in Table 2. We see that AHMC has comparable performance for all models except for HPR, which could be the result of the difference in the implementation of the probability mass function for the Poisson distribution. Also note that AHMC scales better for LBA models while Stan scales better for the rest of the models. This is likely due to the fact that the LBA model relies on a manually implemented distribution which is highly optimized by Julia’s compiler.

5. Conclusion

We have introduced AdvancedHMC.jl, a new Julia package that provides robust, modular and efficient implementations of advanced HMC algorithms. We highlight the modularity of the package and compare the implemented NUTS with Stan, both statistically and computationally. Overall, we hope that AHMC can serve as a robust baseline for future research on HMC algorithms and can enable the development of new PPLs which take advantage of the decoupling of inference algorithms from the actual modelling language.
Acknowledgments

We would like to thank the developers of MCMCBenchmarks.jl, Rob Goedman and Christopher Fisher. With this package we ran all benchmarks and generated all plots presented in this paper. Since AdvancedHMC is an open source project, we have received many contributions from volunteers around the world. We would like to thank all contributors for their valuable work.

References


Appendix A. Full descriptions of models used for evaluations

Below are the full descriptions of five models used for evaluations.

**Gaussian Model (Gaussian)** is a simple two parameter Gaussian distribution.

\[
\begin{align*}
\mu & \sim \mathcal{N}(0, 1) \\
\sigma & \sim \text{Truncated}(\text{Cauchy}(0, 5), 0, \infty) \\
y_n & \sim \mathcal{N}(\mu, \sigma) \quad (n = 1, \ldots, N)
\end{align*}
\]

**Signal Detection Model (SDT)** is a model used in psychophysics and signal processing, which decomposes performance in terms of discriminability and bias (Green et al., 1966).

\[
\begin{align*}
d & \sim \mathcal{N}(0, \frac{1}{\sqrt{2}}) \\
c & \sim \mathcal{N}(0, \frac{1}{\sqrt{2}}) \\
x & \sim \text{SDT}(d, c)
\end{align*}
\]

**Linear Regression Model (LR)** is a linear regression with truncated Cauchy prior on the weights.

\[
\begin{align*}
B_d & \sim \mathcal{N}(0, 10) \\
\sigma & \sim \text{Truncated}(\text{Cauchy}(0, 5), 0, \infty) \\
y_n & \sim \mathcal{N}(\mu_n, \sigma)
\end{align*}
\]

where \( \mu = B_0 + B^T X \), \( d = 1, \ldots, D \) and \( n = 1, \ldots, N \).

**Hierarchical Poisson Regression (HPR)**

\[
\begin{align*}
a_0 & \sim \mathcal{N}(0, 10) \\
a_1 & \sim \mathcal{N}(0, 1) \\
b_\sigma & \sim \text{Truncated}(\text{Cauchy}(0, 1), 0, \infty) \\
b_d & \sim \mathcal{N}(0, b_\sigma) \\
y_n & \sim \text{Poisson}(\log \lambda_n)
\end{align*}
\]

where \( \log \lambda_n = a_0 + b_{z_n} + a_1 x_n \), \( d = 1, \ldots, N_b \) and \( n = 1, \ldots, N \).

**Linear Ballistic Accumulator (LBA)** is a cognitive model of perception and simple decision making (Brown and Heathcote, 2008).

\[
\begin{align*}
\tau & \sim \text{Truncated}(\mathcal{N}(0.4, 0.1), 0, mn) \\
A & \sim \text{Truncated}(\mathcal{N}(0.8, 0.4), 0, \infty) \\
k & \sim \text{Truncated}(\mathcal{N}(0.2, 0.3), 0, \infty) \\
v_d & \sim \text{Truncated}(\mathcal{N}(0, 3), 0, \infty) \\
x_n & \sim \text{LBA}(\nu, \tau, A, k)
\end{align*}
\]

where \( mn = \min_i x_{i,2} \), \( d = 1, \ldots, N_c \) and \( n = 1, \ldots, N \).
Appendix B. Statistical property of simulated trajectories

<table>
<thead>
<tr>
<th>N</th>
<th>ESS</th>
<th>μ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stan</td>
<td>10</td>
<td>513.163</td>
<td>466.577</td>
</tr>
<tr>
<td>AHMC</td>
<td>10</td>
<td>563.535</td>
<td>447.722</td>
</tr>
<tr>
<td>Stan</td>
<td>100</td>
<td>786.531</td>
<td>782.231</td>
</tr>
<tr>
<td>AHMC</td>
<td>100</td>
<td>786.531</td>
<td>796.628</td>
</tr>
<tr>
<td>Stan</td>
<td>1000</td>
<td>864.010</td>
<td>876.660</td>
</tr>
<tr>
<td>AHMC</td>
<td>1000</td>
<td>832.255</td>
<td>844.452</td>
</tr>
</tbody>
</table>

Figure 2: Gaussian (50 runs); left to right: step size, tree depth, ESS

<table>
<thead>
<tr>
<th>N</th>
<th>d</th>
<th>ESS</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stan</td>
<td>10</td>
<td>710.762</td>
<td>703.327</td>
</tr>
<tr>
<td>AHMC</td>
<td>10</td>
<td>882.236</td>
<td>815.929</td>
</tr>
<tr>
<td>Stan</td>
<td>100</td>
<td>820.741</td>
<td>823.152</td>
</tr>
<tr>
<td>AHMC</td>
<td>100</td>
<td>814.308</td>
<td>846.357</td>
</tr>
<tr>
<td>Stan</td>
<td>1000</td>
<td>844.478</td>
<td>872.961</td>
</tr>
<tr>
<td>AHMC</td>
<td>1000</td>
<td>829.792</td>
<td>859.018</td>
</tr>
</tbody>
</table>

Figure 3: SDT (100 runs); left to right: step size, tree depth, ESS

<table>
<thead>
<tr>
<th>N</th>
<th>a₀</th>
<th>ESS</th>
<th>a₁</th>
<th>b₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stan</td>
<td>10</td>
<td>221.485</td>
<td>215.013</td>
<td>266.900</td>
</tr>
<tr>
<td>AHMC</td>
<td>10</td>
<td>216.491</td>
<td>214.549</td>
<td>258.638</td>
</tr>
<tr>
<td>Stan</td>
<td>20</td>
<td>208.286</td>
<td>207.041</td>
<td>241.080</td>
</tr>
<tr>
<td>AHMC</td>
<td>20</td>
<td>206.458</td>
<td>206.469</td>
<td>236.546</td>
</tr>
<tr>
<td>Stan</td>
<td>50</td>
<td>172.484</td>
<td>172.982</td>
<td>216.586</td>
</tr>
<tr>
<td>AHMC</td>
<td>50</td>
<td>200.755</td>
<td>201.548</td>
<td>247.384</td>
</tr>
</tbody>
</table>

Figure 4: HPR (25 runs); left to right: step size, tree depth, ESS (of some variables)

<table>
<thead>
<tr>
<th>N</th>
<th>τ</th>
<th>A</th>
<th>ESS</th>
<th>ν₁</th>
<th>ν₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stan</td>
<td>10</td>
<td>226.463</td>
<td>252.656</td>
<td>305.614</td>
<td>276.557</td>
</tr>
<tr>
<td>AHMC</td>
<td>10</td>
<td>340.722</td>
<td>304.523</td>
<td>357.610</td>
<td>336.357</td>
</tr>
<tr>
<td>Stan</td>
<td>50</td>
<td>212.838</td>
<td>238.003</td>
<td>24.089</td>
<td>232.667</td>
</tr>
<tr>
<td>AHMC</td>
<td>50</td>
<td>248.249</td>
<td>238.979</td>
<td>248.331</td>
<td>255.421</td>
</tr>
<tr>
<td>Stan</td>
<td>200</td>
<td>244.926</td>
<td>264.967</td>
<td>268.793</td>
<td>270.36</td>
</tr>
<tr>
<td>AHMC</td>
<td>200</td>
<td>256.638</td>
<td>263.098</td>
<td>270.978</td>
<td>266.769</td>
</tr>
</tbody>
</table>

Figure 5: LBA (50 runs); left to right: step size, tree depth, ESS (of some variables)
Appendix C. Turing code for the linear regression model

Below is the code snippet of running NUTS using Turing for the LR model.

```julia
@model LR(x, y, Nd, Nc) = begin
    B ~ MvNormal(zeros(Nc), 10)
    B0 ~ Normal(0, 10)
    sigma ~ Truncated(Cauchy(0, 5), 0, Inf)
    mu = B0 .+ x * B
    y ~ MvNormal(mu, sigma)
end
x, y, Nd, Nc = ... # load data
chain = sample(LR(x, y, Nd, Nc), NUTS(2_000, 1_000, 0.8))
```