An Adaptive MCMC Scheme for Setting Trajectory Lengths in Hamiltonian Monte Carlo: Supplemental Materials

Matthew D. Hoffman
Google Research

Alexey Radul
Google Research

Pavel Sountsov
Google Research

1 Derivation of autocorrelation of second moment in the Gaussian case

In this section, we derive equation 5 from the main text, describing the lag-1 autocorrelation of the second moment of a scalar θ under HMC targeting a Gaussian with mean 0 and scale σ:

$$\frac{\mathbb{E}[((\theta')^2 - \mathbb{E}[\theta^2])(\theta^2 - \mathbb{E}[\theta^2])]}{\mathbb{E}[(\theta^2 - \mathbb{E}[\theta^2])^2)]} = \frac{1}{2}(1 + \cos(2t/\sigma)). \tag{1}$$

To reduce clutter, we omit subscripts and focus on the scalar case. Let the position θ be a sample from $\mathcal{N}(0, \sigma)$, the momentum r be a sample from $\mathcal{N}(0, 1)$ and θ', r' be the result of evolving the initial state θ, r for some time t according to the continuous-time Hamiltonian dynamics

$$\frac{\partial \theta}{\partial t} = m; \quad \frac{\partial r}{\partial t} = -\frac{\theta}{\sigma^2}. \tag{2}$$

Then marginalizing out θ, $\theta' \sim \mathcal{N}(0, \sigma)$, since HMC with exact simulation leaves the distribution of θ invariant and has an acceptance probability of 1, and in particular

$$\mathbb{E}[\theta^2] = \mathbb{E}[(\theta')^2] = \sigma^2. \tag{3}$$

Recall that

$$\begin{aligned} \theta = a\sigma \sin(\phi); \quad r = a\cos(\phi); & \qquad a^2 = \theta^2 + r^2; \\ \theta' = a\sigma \sin(\phi + t/\sigma); \quad r' = a\cos(\phi + t/\sigma); & \qquad \phi = \tan^{-1}(\tfrac{\theta}{\sigma r}), \end{aligned} \tag{4}$$

and that at equilibrium the phase-space angle ϕ and magnitude a are independent with $\phi \sim \text{Uniform}(0, 2\pi)$ and $a^2 \sim \text{Exponential}(2)$, so $\mathbb{E}[a^4] = 8$.

The mean and variance of θ^2 at equilibrium are σ^2 and $2\sigma^4$, so the lag-1 autocorrelation of a scalar Gaussian θ^2 is

$$\begin{aligned} \frac{1}{2\sigma^4}\mathbb{E}[((\theta')^2 - \sigma^2)(\theta^2 - \sigma^2)] &= \frac{1}{2\sigma^4}\mathbb{E}[(\theta')^2\theta^2 - \sigma^2(\theta')^2 - \sigma^2\theta^2 + \sigma^4] \\ &= \frac{1}{2\sigma^4}\left(\mathbb{E}[(\theta')^2\theta^2] - \sigma^2\mathbb{E}[(\theta')^2] - \sigma^2\mathbb{E}[\theta^2] + \sigma^4\right) \\ &= \frac{1}{2\sigma^4}\left(\mathbb{E}[(\theta')^2\theta^2] - \sigma^4\right) \\ &= \frac{1}{2\sigma^4}\mathbb{E}[(\theta')^2\theta^2] - \frac{1}{2} \\ &= \frac{1}{2}\mathbb{E}[(a\sin(\phi + t/\sigma))^2 (a\sin(\phi))^2] - \frac{1}{2} \\ &= \frac{1}{2}\mathbb{E}[a^4 \sin^2(\phi + t/\sigma)\sin^2(\phi)] - \frac{1}{2} \\ &= \frac{1}{2}\mathbb{E}[a^4]\mathbb{E}[\sin^2(\phi + t/\sigma)\sin^2(\phi)] - \frac{1}{2} \\ &= 4\mathbb{E}\left[\frac{1}{2}(1 - \cos(2\phi + 2t/\sigma))\frac{1}{2}(1 - \cos(2\phi))\right] - \frac{1}{2} \\ &= \mathbb{E}[1 - \cos(2\phi + 2t/\sigma) - \cos(2\phi) + \cos(2\phi)\cos(2\phi + 2t/\sigma)] - \frac{1}{2} \\ &= \mathbb{E}[1 - \cos(2\phi + 2t/\sigma) - \cos(2\phi) + \cos(2t/\sigma) + \cos(4\phi + 2t/\sigma)] - \frac{1}{2} \\ &= \frac{1}{2}(1 + \cos(2t/\sigma)), \end{aligned} \tag{5}$$

where we use the fact that, if $\phi \sim \text{Uniform}(0, 2\pi)$,

$$\mathbb{E}[\cos(n\phi + \delta)] = 0 \tag{6}$$

for any δ and integer n. We also use the standard trigonometric identities

$$\sin^2(x) = \frac{1}{2}(1 - \cos(2x)); \quad \cos(x)\cos(x + y) = \frac{1}{2}(\cos(y) + \cos(2x + y)). \tag{7}$$

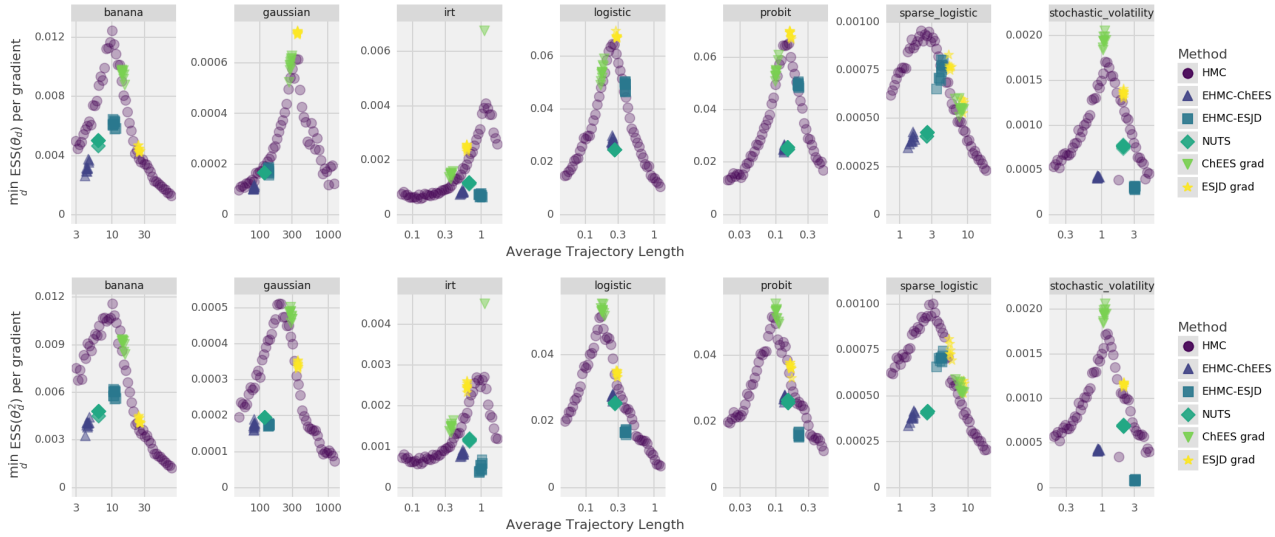

Figure 1: Effective sample size (ESS) per gradient evaluation as a function of average number of leapfrog steps for jittered HMC with fixed average trajectory length, HMC tuned by maximizing ChEES and ESJD, NUTS, and empirical HMC (EHMC) tuned to maximize both ESJD (as originally proposed by Wu et al. (2018)) and ChEES. Each point is a single run of 100 chains, each plot within a row represents a different target distribution. For NUTS, the average trajectory length is based on the number of leapfrog steps separating the initial and final state, which is about half the total number of leapfrog steps taken on average since NUTS must "waste" some leapfrog steps to satisfy detailed balance. ESS is evaluated in terms of the estimator of the first moment ($\mathbb{E}[\theta]$; top plots) and second moment ($\mathbb{E}[\theta^2]$; bottom plots), and for each statistic we report the minimum across dimensions of the median ESS across chains. Insofar as a good sampler should yield good estimators of any posterior expectation, when ESS for the first and second moment are different we consider the smaller of the two to be more the more meaningful summary.

2 Effective Sample Size Plots

Figure 1 summarizes the relationship between effective sample size (ESS) per gradient (computed with respect to first and second moment estimates) and trajectory length for a variety of algorithms. ESS/gradient is a computational-substrate-independent measure of sampling efficiency; higher is better.

The purple circles ("HMC") were obtained by running a grid search over trajectory lengths, which would be very expensive in practice. Trajectory lengths that yield optimal ESS/gradient in terms of first moments (top plots) can yield poor ESS/gradient in terms of second moments (bottom plots), particularly for the Gaussian, logistic regression, and probit regression targets.

All other methods evaluated tune their trajectory lengths automatically, and all of them consistently choose average trajectory lengths that are within about factor of three of the optimal trajectory length. However, being off by a factor of two can be expensive, and different methods have different amounts of overhead: NUTS must spend extra leapfrog steps to satisfy detailed balance, and empirical HMC ("EHMC-ESJD"; Wu et al., 2018) must often spend extra leapfrog steps during warmup to estimate the empirical distribution of ESJD-maximizing trajectory lengths.

ChEES-HMC ("ChEES grad") often finds trajectory lengths very close to those that grid-search-tuned HMC would have; for the item-response theory and sparse logistic regression targets, it still produces ESS/gradient results within a factor of two of optimal. Maximizing expected squared jumped distance ("ESJD grad") instead of ChEES usually yields longer trajectory lengths, which are usually suboptimal.

We also evaluated EHMC maximizing ChEES instead of ESJD, which sometimes (but not always) yielded an improvement.

3 Halton Versus Uniform Jitter Ablation Plots

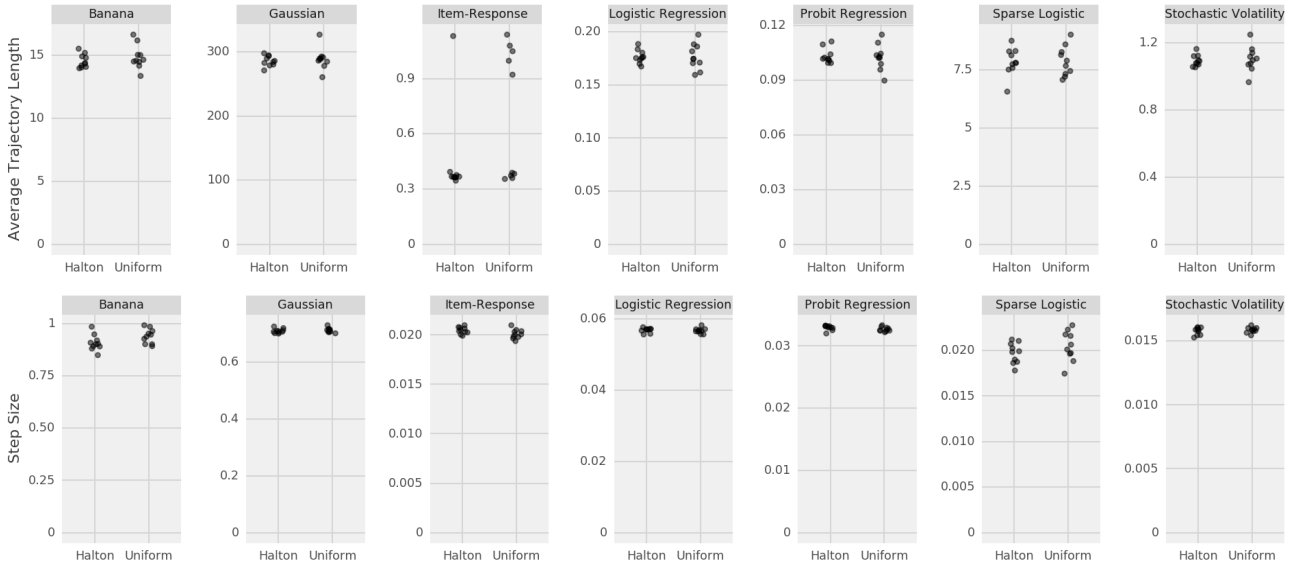

Figure 2: Trajectory length and step size obtained by ChEES-HMC and dual averaging, jittering the trajectory length between 0 and T using uniform random numbers or the Halton sequence. Using the Halton sequence instead of uniform pseudorandom numbers to jitter trajectory lengths results in less variability in the tuned trajectory length, dramatically so in the item-response-theory posterior target.

Figure 2 shows the variability of trajectory lengths and step sizes found by ChEES-HMC when jittering number of leapfrog steps using a Halton sequence or uniform random numbers. Using Halton sequences instead of uniform random numbers consistently reduces the variability of the trajectory length chosen by ChEES-HMC (sometimes drastically, as in the item-response theory target).

The step-size tuning procedure exhibits relatively little variation under either kind of jitter. This indifference to the type of jitter is consistent with the claim that, for long-enough simulation times, the simulation error of the leapfrog integrator depends only weakly on simulation time (Hairer et al., 2006).

4 Adaptation Speed Ablation

Figure 3 examines the effect of the Adam adaptation speed (often called α) on the stability of the tuned trajectory length. The results in the paper use $\alpha = 0.025$, which is a relatively conservative choice—most models we look at are not very sensitive to this parameter. The item-response theory model is an exception; using a larger adaptation speed of 0.0375 would have led to a longer average trajectory length, which would have yielded better results (see figure 1).

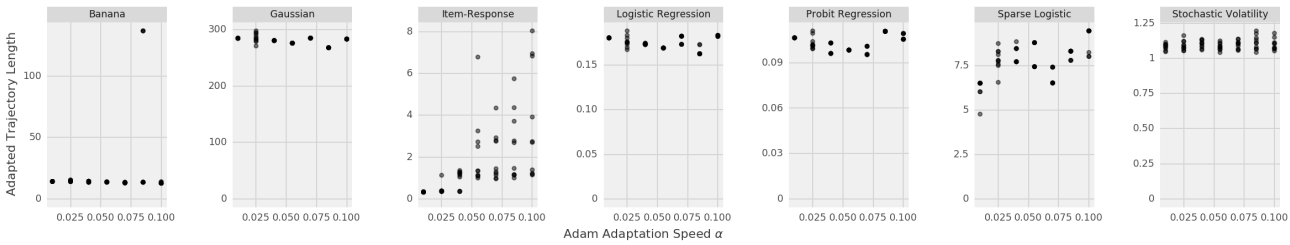

Figure 3: ChEES-HMC adapted trajectory length as a function of adaptation speed. Each point is an independent run with 100 chains.

5 Cross-Chain Adaptation Stuck Chain Example

Figure 4 shows an example of chains getting stuck during cross-chain step-size adaptation tuning arithmetic instead of harmonic mean targeting the sparse logistic regression posterior. Using the arithmetic mean, many chains are stuck; the harmonic mean makes sure that all chains move.

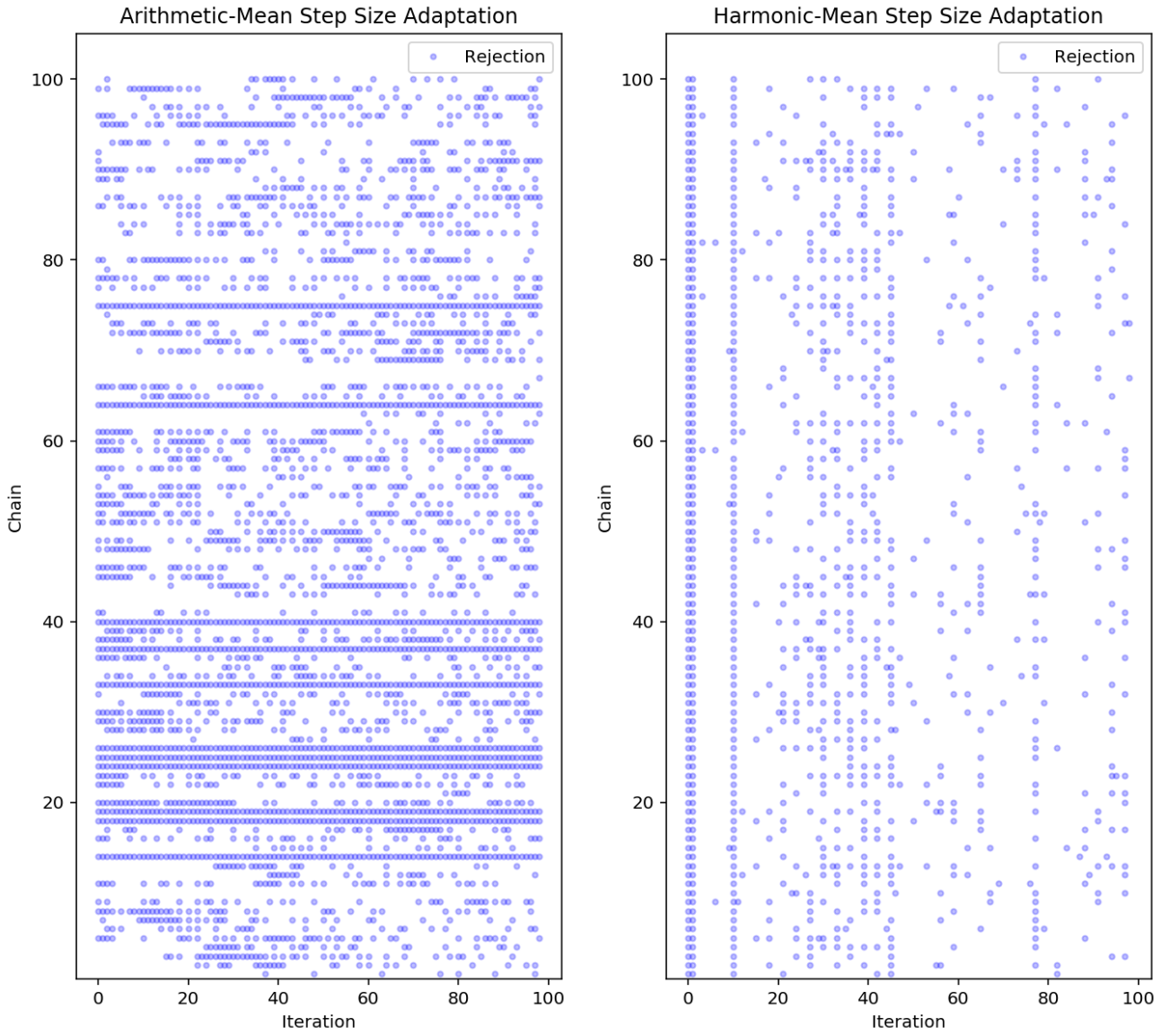

Figure 4: Rejections by chain id and iteration number.

References

Hairer, E., Lubich, C., and Wanner, G. *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

Wu, C., Stoehr, J., and Robert, C. P. Faster hamiltonian monte carlo by learning leapfrog scale. October 2018.