RSAP: an adaptive Metropolis algorithm with rejection-based Gaussian proposal-scaling for fast convergence in multimodal parameter spaces *

Graham West †, John Wallin ‡, and Zachariah Sinkala §

Abstract. Performing MCMC parameter estimation on complex, nonlinear mathematical models can quickly lead to endless searching through highly multimodal parameter spaces. If one does not know the optimal proposal distribution, the burn-in period can take an excessively long time, due to the Markov chain inefficiently mixing and becoming trapped near a suboptimal mode. Also, if the model of interest is very computationally expensive, one would prefer to evaluate it as few times as possible. With these challenges in mind, we present in this paper a new adaptive Metropolis method called RSAP (rejection-scaled adaptive proposal) which can grow and shrink the scale of the Gaussian proposal width (independently for each parameter) based on the number of consecutive rejections the chain has encountered. The method is flexible and robust enough to handle parameter spaces that are both high-dimensional and multimodal. Other advantages include not having to locate a near-optimal with a different method beforehand, as well as requiring almost zero computational or storage overhead from standard Metropolis. It also solves the issue of the unknown optimal proposal width by adjusting it over time. We will provide a proof of its ergodicity and conclude with comprehensive computational results implementing this method on several benchmark optimization functions, comparing its convergence to that of the standard Metropolis method and a modified version of Adaptive Metropolis [9].

Key words. Markov chain Monte Carlo, adaptive MCMC methods, parameter estimation, ergodicity, proposal-scaling

AMS subject classifications. 62M05 (markovian estimation), 62M09 (non-markovian estimation)

1. Introduction.

Markov chain Monte Carlo (MCMC) methods provide one of the most widely-used and easily accessible solution techniques to the problem of model parameter estimation (PE). As the name suggests, MCMC algorithms use Markov chains to sample the state space of the model being studied, thus obtaining distributions for all of its relevant parameters. With an ever-growing number of MCMC method types—from interacting forests of chains to delayed rejection techniques which approximate a full model evaluation—a suitable method can be found for virtually every problem.

This paper is outlined in the following way. First, we open with a review of the original Metropolis method, as well as a discussion on how kernel mixing and adaptive methods are able to increase its performance. Next, we will discuss our own adaptive kernel mixing method RSAP and the motivating factors for using it, as well as provide a proof of its ergodicity. Lastly, we will present and interpret results of experiments performed on our method using common optimization benchmark functions (including Gaussian and Ackley functions as well

*Submitted to the editors DATE.

†PhD Program in Computational Science, Middle Tennessee State University, Murfreesboro, TN (gtw2i@mtmail.mtsu.edu).

‡PhD Program in Computational Science and Department of Physics & Astronomy, Middle Tennessee State University, Murfreesboro, TN (jwallin@mtsu.edu).

§Department of Mathematical Sciences and PhD Program in Computational Science, Middle Tennessee State University, Murfreesboro, TN (zsinkala@mtsu.edu).

as a flattened bi-modal function), going into detail in comparing our method's performance with that of standard Metropolis [3] as well as Adaptive Metropolis [9] over a range of proposal widths.

1.1. Metropolis. One of the most well-known and widely-used MCMC techniques is the original Metropolis method. First developed by Metropolis et al. [3] and later generalized by Hastings [2], this simple method has been the foundation for countless new and improved versions that have continued to build on the basic principles found within their predecessor.

Given data that we wish for the model to fit, the Metropolis method attempts to sample a distribution over the model parameters called the *posterior*. This distribution will have a peak at the values of the model parameters which best approximate the data. However, we do not know the posterior since we cannot directly invert the parameter-data relationship (and since to do so would be to know the solution to the problem). Therefore, we must express it in terms of distributions which we do know. From Bayes's Theorem, we have the following proportionality relation (given data y and model parameters θ)

$$\pi(\theta|y) \propto \ell(y|\theta)p(\theta)$$

where π is the *posterior* distribution, ℓ is the *likelihood* distribution, and p is the *prior* distribution. The likelihood $\ell(y|\theta)$ tells us the probability of the data given the parameters. Therefore, we can easily calculate its value via a single model evaluation with the given parameters θ (referred to as the state). In addition, we may also have prior information about what parameter values are most likely to be realized in the actual system (for example, if the parameters can only take on values in a finite range or tend to be concentrated about a central mean). This information is incorporated via the prior distribution p. Together, the likelihood and prior help us to reconstruct the posterior distribution (up to a scale factor which does not impact the method) so that the inverse problem may be solved.

Now, concerning the specifics of how the Metropolis method successfully samples the posterior, suppose we're given some data y and an initial state θ^0 (for this paper, superscript indices will refer to the time step), which is a reasonable guess for the proper values of the parameters. The method generates new candidate states and probabilistically accepts them based on their relative improvement from the current state, using the Bayesian approximation of the posterior as a metric. Candidate state generation is performed via sampling of the *proposal* distribution $q(\theta'|\theta^n)$, defined as the probability that a new state θ' will be selected as a candidate for acceptance given the current state θ^n (a prime denotes a candidate state). Whether acceptance or rejection is performed is determined by the *acceptance probability* $\alpha(\theta'|\theta^n)$, defined as the probability that the candidate state θ' will be accepted given the current step θ^n. Over time, the distribution created from the samples will conform to the posterior, at which point the method is said to have *converged*. Below is a pseudocode implementation of Metropolis (Algorithm 1.1) which runs for N_{step} steps.

The Metropolis method has several useful properties which contribute its effectiveness. First—interpreting the method similar to an annealing-type optimization method—it can both naively hill-climb and tactically accept poorer states to escape suboptimal local modes. When testing a new candidate state for acceptance, if it has an equal or higher posterior value than the current state, then $\alpha = 1$ and acceptance is guaranteed (hill-climbing). On the other hand,

Algorithm 1.1 The Metropolis method

1: Initialize θ^0
2: **for** $n = 1$ **to** N_{step} **do**
3: Generate $\theta' \sim q(\theta'|\theta^{n-1})$
4: Compute $\alpha(\theta'|\theta^{n-1}) = \min\left(1, \dfrac{\pi(\theta'|y)}{\pi(\theta^{n-1}|y)}\right)$
$= \min\left(1, \dfrac{\ell(y|\theta')p(\theta')}{\ell(y|\theta^{n-1})p(\theta^{n-1})}\right)$
5: Set $\theta^n = \theta'$ with probability α, else $\theta^n = \theta^{n-1}$
6: **end for**

if the posterior is lower, then $1 > \alpha \geq 0$ and the poorer state still has a probability of being accepted. Though poorer states can be accepted, α decreases as the candidate's posterior becomes lower, so very poor candidates will usually be rejected. Counter-intuitively, this ability to accept poorer states is required to guarantee the eventual convergence of Metropolis (as evidenced by the appearance of the acceptance probability in the proof in the following paragraph).

This leads us to the second useful property of Metropolis. As long as the proposal distribution endows the Markov chain with a few key properties (irreducibility, aperiodicity, and non-transcience)—which hold for nearly all reasonable proposals—the distribution will always converge to the posterior, the stationary distribution of the chain formed by Metropolis. Assuming the three conditions mentioned above hold for the chain, we can easily show that the posterior distribution is the stationary distribution of the Metropolis method (proof taken from Gelman [4] and included for convenience of the reader). Say the algorithm is at the n-th step and that $\theta^n = a$, where a was drawn from the posterior. Now, suppose we propose the candidate $\theta^{n+1} = b$ such that $\pi(b) \geq \pi(a)$. Then,

$$P(\theta^n = a, \theta^{n+1} = b) = \min\left(1, \frac{\pi(b)}{\pi(a)}\right) q(b|a)\pi(a) = q(b|a)\pi(a)$$

On the other hand, suppose $\theta^n = b$ and we attempt a transition to the candidate $\theta^{n+1} = a$. Then,

$$\begin{aligned} P(\theta^n = b, \theta^{n+1} = a) &= \min\left(1, \frac{\pi(a)}{\pi(b)}\right) q(a|b)\pi(b) \\ &= \frac{\pi(a)}{\pi(b)} q(a|b)\pi(b) = q(a|b)\pi(a) = q(b|a)\pi(a) \end{aligned}$$

Therefore, $P(\theta^n = a, \theta^{n+1} = b) = P(\theta^n = b, \theta^{n+1} = a)$, which is the condition for stationarity.

Note that since the stationary distribution—being simply the posterior—is independent of the proposal used, and thus holds for all valid proposals. However, though the method is guaranteed to always converge to π, it may take a large number of steps for convergence to

occur. Much of the algorithm's utility rests on the choice of a good proposal q. If one chooses too thin of a distribution, the method may have difficulty escaping suboptimal modes within an acceptable amount of time. On the other hand, too wide a distribution and the method might jump past true optimal modes or begin to oscillate about them, taking a large number of steps to sink into them.

With this in mind, the key question is, "What is the optimal proposal to use for problem X?" As should be expected, to know the answer to this question is essentially to know the solution to problem X. Therefore, certain assumptions and compromises must be made when designing proposals. Fortunately, as shall be seen in the following sections, this problem of the unknown proposal can be largely alleviated through clever means such as kernel mixing and adaptive proposals.

1.2. Kernel mixing and composition. Kernel mixing and composition are techniques where the transition probabilities (or kernel) used at each step in the chain may be different. Specifically, *composition* refers to a method where the kernel is chosen at each step through some deterministic means. Consider, for example, the simple case we have two kernels P_0, P_1 and simply alternate between them. This produces the n-step kernel

$$\prod_{j=0}^{n-1} P_{j \bmod 2}$$

where it is understood that we are using the convention commonly found in Markov chain literature of right multiplication by the kernel.

On the other hand, techniques where the kernels used at each step are determined by some random process are termed *mixing* methods. Consider another example where we have two kernels P_0, P_1, but instead of strictly alternating, we have a probability 0.5 of choosing either at any given step. Then, the kernel used at each step is simply $0.5P_0 + 0.5P_1$—the linear combination of the kernels and their respective probabilities—and the n-step kernel is $\left(0.5P_0 + 0.5P_1\right)^n$. Upon an actual run of this method, the true n-step kernel used is

$$\prod_{j=0}^{n-1} P_{i_j}$$

where the i_j are determined by performing some particular stochastic process. However, when run for a large number of steps, the second kernel will converge to the first. As shall be seen in Section 2, our method uses kernel mixing by randomly selecting between numerous Gaussian proposals of varying widths.

1.3. Adaptive MCMC. Another technique for improving the performance of MCMC is to use so-called *adaptive* methods which allow the kernel to adapt to the state space as more and more samples are obtained. The chains produced by these methods are referred to as *inhomogeneous* chains since the kernel is not constant over time. For many models with nonlinear interactions between the parameters, neighboring regions in state space can have completely different posterior profiles, causing what would be an optimal proposal in one region to perform poorly in another. There may be numerous local modes clustered together

or large plateaus. These effects combined with a lack of prior knowledge of which proposal should be chosen can inevitably lead to long run times. Thus, in these cases, one can justify the use of methods which can tune themselves based on the changing geometry of the state space.

Though adaptive methods can dramatically increase computational performance and efficiency, they often lose some of the desired analytical properties that their more simple counterparts have. For example, if the evolution of the kernel depends on the samples from multiple previous steps, the chain is no longer *Markovian*. Also, in general, one cannot prove ergodicity without the assumption of several special conditions which may be difficult to verify in practice. Fortunately, even if one cannot verify convergence properties for a specific adaptive MCMC method, it may still be used to dramatically shorten the burn-in time as the chain searches for the target distribution. Despite the difficulties discussed, we were actually able to prove our own method's ergodicity through simple (yet notationally dense) means thanks to Theorem 5 from a paper by Roberts and Rosenthal [8]. See this resource for a very thorough analysis on conditions for the convergence of adaptive MCMC methods.

Two excellent examples of adaptive MCMC methods come from Haario, Saksman, and Tamminen in the form of the *Adaptive Proposal* (AP) [5] and *Adaptive Metropolis* (AM) [6] methods. While both methods are based on the same principle of adaptation, AM retains ergodicity while AP does not. The core principle shared by the two is the use of a continuously adapting Gaussian distribution as the proposal for the standard Metropolis method. After beginning with a Metropolis-only phase to acquire some initial samples, the proposal used is

$$q^n(\theta'|\theta^n) = N\left(\theta^n, \left(\frac{2.38^2}{M}\right)\Sigma^n + \varepsilon I\right) \tag{1.1}$$

where M is the dimension of the space (i.e., number of parameters), Σ^n is the covariance matrix computed from samples generated over past steps, and εI is a small factor multiplied by an $M \times M$ identity matrix (included so that degeneracy of Σ^n does not cause the method to collapse). This allows the Gaussian to continually re-scale and re-orient itself as it encounters different regions of state space. Now, the only difference between AP and AM is that AP uses a fixed number of samples to calculate Σ^n while AM uses the entire history of the chain. Thus, AP's adaptation does not vanish since its covariance matrix does not converge, while AM's does (due to the central limit theorem).

Equation (1.1) can be modified in several ways. For example, one is not restricted to use of ϵI. A more appropriate fixed covariance matrix may replace it. Also, one can employ kernel mixing (rather than mixing just the covariance matrices). A good example of this modified AM is the version developed by Roberts and Rosenthal [9]

$$q^n(\theta'|\theta^n) = \beta N\left(\theta^n, \left(\frac{2.38^2}{M}\right)\Sigma^n\right) + (1-\beta)N\left(\theta^n, \hat{\Sigma}\right)$$

where $\hat{\Sigma}$ is a fixed covariance matrix and $0 \leq \beta \leq 1$ (note that our convention is different from theirs in that we have swapped the β and $(1-\beta)$ terms). Following their lead, the version of

AM that we implemented for comparison with RSAP has the following form

$$q^n(\theta'|\theta^n) = N\left(\theta^n, \beta\left(\frac{2.38^2}{M}\right)\Sigma^n + (1-\beta)\hat{\Sigma}\right)$$

though, in our numerical tests, it was always such that $\hat{\Sigma} \propto I$ (as in Equation (1.1)).

2. Rejection-scaled adaptive proposal (RSAP). We now move on to present our own contribution to the field of adaptive MCMC: the rejection-scaled adaptive proposal method (RSAP). Based on the original Metropolis method, this method uses very simple means to scale the entries of the Gaussian proposal's diagonal covariance matrix (i.e., the proposal widths in the direction of each model parameter) to more locally optimal values. We will begin with a discussion of our motivation for developing the method, the challenges it addresses, and some advantages it has over standard Metropolis and some other adaptive methods. We will then move on to a detailed description of the method on a step-by-step level and end with our ergodicity proof.

2.1. Motivation and advantages. There are several advantages RSAP has over standard Metropolis or other adaptive methods. First, as we have discussed, it isn't possible in general to know a priori what the optimal proposal width should be for a given problem. Even if one chooses the optimal proposal width, this by no means guarantees that certain regions of state space might not be more efficiently traversed with a different proposal width. These issues are resolved by the ability of RSAP to both grow and shrink the proposal width at a set amplitude and rate, allowing the chain to escape local modes in fewer steps and traverse flatter regions more quickly. In addition, the adaptation can swap between growth and shrinkage of the width across consecutive time steps, thus allowing for discontinuous changes in the adaptive scaling parameters. This allows the method to not be delayed as long as it might otherwise be when finding a more optimal width.

A second advantage is that our method does not require an initial greedy search procedure to find a local or global mode of the posterior. Due to its powerful adaptability, RSAP can act as its own greedy search or burn in method.

Yet another advantage that RSAP has over other more complicated adaptive methods is its trivial implementation and the minimal overhead it requires. Both the extra storage and computing time required for our method compared to standard Metropolis are negligible, since only several additional tuning parameters, counters, and scale factors must be stored and the adaptive scaling functions are computationally trivial to evaluate. Because of this, there is essentially no reason not to use RSAP as opposed to standard Metropolis.

2.2. The method. The key to the RSAP method is that it mixes a fixed-width Gaussian with two adaptable wide and thin Gaussians (one of each) into a single proposal. The thin option is included so that the chain can squeeze into tight optimal modes and the wide option so that the chain can both escape suboptimal modes and traverse state space quickly. The fixed option is included so that we always have a reasonable control to use in case the others adapt "too much" (which is unlikely given our constrants on the scaling). With this reasoning in mind, the proposal for a single model parameter has the form of the linear combination

$$q^n(\theta'|\theta^n) = p_t^n N\left(\theta^n, (\sigma_t^n)^2\right) + p_f^n N\left(\theta^n, (\sigma_f)^2\right) + p_w^n N\left(\theta^n, (\sigma_w^n)^2\right) \tag{2.1}$$

where $\sigma_t^n \leq \sigma_f \leq \sigma_w^n$ are the proposal widths and $0 \leq p_t^n, p_f^n, p_w^n \leq 1$ are the mixing probabilities. Every parameter uses separately adapting wide and thin proposal widths—allowing for

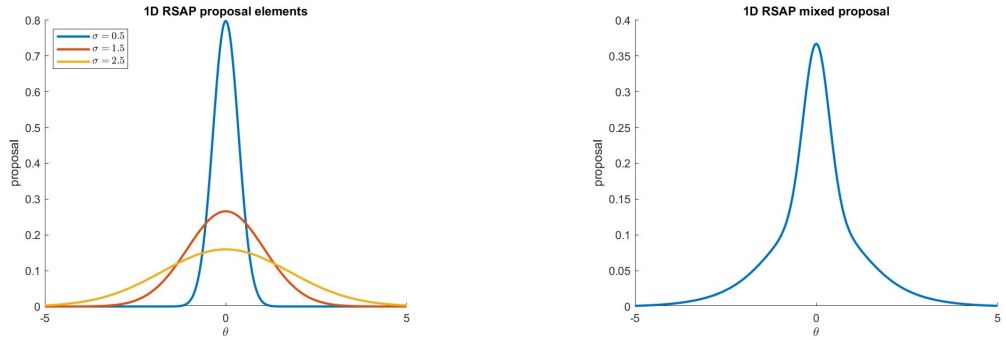

Figure 1. *RSAP proposal (Equation (2.1)) with* $\sigma_t = 0.5, \sigma_f = 1.5, \sigma_w = 2.5, p_t = p_w = p_f = 1/3$.

independent adaptation of each—as well as its own fixed width (since the different parameters likely have different units and/or scales).

The values of σ_t^n, σ_w^n adapt based on the number of recent consecutive rejections the chain has encountered, with their default values set to σ_f. If the candidate produced by the method is rejected, we choose from the three kernels based on the defined mixing probabilities (which we will discuss in detail below). If the fixed kernel is chosen, then no adaptation is performed. However, if either the thin or wide kernels are used, then we increment one of the counters k_t^n, k_w^n and compute the appropriate scale factor

$$\begin{aligned} A_t(k_t^n) &= 1 - (1 - \hat{A}_t)(1 - \exp(-r_t k_t^n)) \\ A_w(k_w^n) &= 1 - (1 - \hat{A}_w)(1 - \exp(-r_w k_w^n)) \end{aligned} \tag{2.2}$$

where $r_w > 0$, $r_t > 0$, $\hat{A}_w > 1$, and $1 > \hat{A}_t > 0$. (These functions were chosen since their initial value is 1, their asymptotic limit is their respective $\hat{A}$ value, and their rate of convergence can be easily controlled via the r constants.) We then set either $\sigma_t^n = A_t(k_t^n)\sigma_f$ or $\sigma_w^n = A_w(k_w^n)\sigma_f$ (depending on which was chosen) and perform a Metropolis step using this proposal width (again remembering that process is done independently for all parameters). On the other hand, whenever a candidate is accepted, the counters are reset to zero so that $\sigma_t^n = \sigma_w^n = \sigma_f$ (for all parameters).

Now, concerning how we choose between the three options, we have the time-dependent mixing probabilities p_t^n, p_f^n, p_w^n. The following formulae govern the specific behavior of these probabilities.

$$p_f^n = \begin{cases} 1/3 & 0 \leq n < n_1 \\ (2 - \cos(\pi(n - n_1)/n_2))/3 & n_1 \leq n < n_1 + n_2 \\ 1 & n_1 + n_2 \leq n < N_{step} \end{cases} \tag{2.3}$$
$$p_t^n = p_w^n = (1 - p_f^n)/2$$

The interval widths n_1, n_2 define three distinct regions for these functions. In the first region—which lasts n_1 steps—they describe a uniform distribution. In practice, n_1 should be set fairly

large w.r.t. the total number of time steps N_{step} so that RSAP has plenty of steps to utilize its adaptive features. The second region is a transition region—which lasts for n_2 steps—where proposal fixing begins to dominate the adaptation. The use of the cosine function here is somewhat arbitrary since any other function that is monotonically increasing on the interval $[n_1, n_1 + n_2]$ would produce similar results. From the definitions of p_t^n, p_w^n, there is an equal probability at each step that either thinning or widening will be chosen. There is no firm guideline for setting the second interval except $n_1 + n_2 \leq N_{step}$. With the last region, we forced eventual convergence to $p_f^n \to 1$, which—as we shall see—implies the ergodicity of RSAP (see Section 2.3).

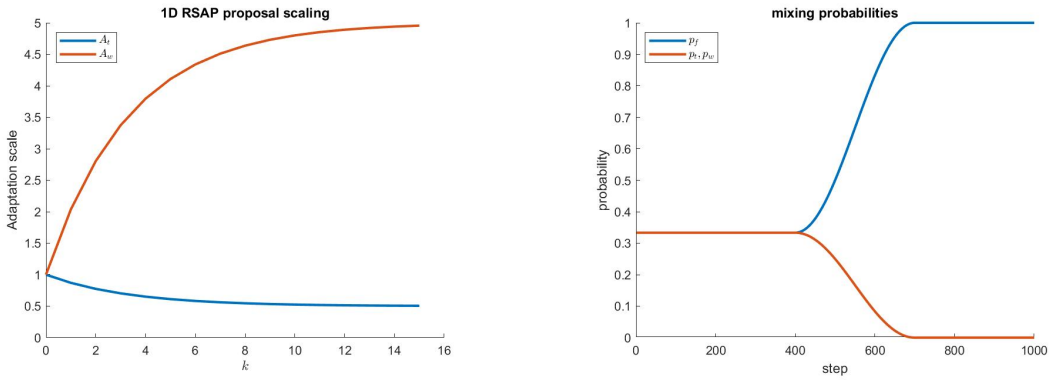

Figure 2. *Left: RSAP proposal-scaling functions* (Equation (2.2)) *vs. adaptation degree* k *with* $r_t, r_w = 0.3, A_t = 0.5, A_w = 5$. *Right: RSAP proposal-mixing probabilities* (Equation (2.3)) *vs. time step with* $n_1 = 400, n_2 = 300$.

To derive the full M-parameter proposal (which we do in Section 2.3 with the help of a great deal of added notation), one has to consider all 3^M possibilities of selecting the M parameters and their three different types of adaptation. It is helpful to do a two-parameter example. Each parameter will have its own set of proposal widths $\sigma_{t,m} < \sigma_{f,m} < \sigma_{w,m}$ $(m = 1, 2)$. Upon adaptation, both parameters must choose which form of adaptation to apply, meaning there are 3^2 possibilities for adaptation. Therefore, defining for the sake of notation $N(\theta^n, \sigma_1^2, \sigma_2^2)$ as a bivariate Gaussian distribution with diagonal covariance matrix given by the diagonal entries (σ_1^2, σ_2^2), we can write the two-parameter proposal as (the superscript n's for σ's omitted for space reasons)

$$
\begin{aligned}
q^n(\theta'|\theta^n) &= \sum_{i\in\{t,f,w\}} \sum_{j\in\{t,f,w\}} p_i^n p_j^n N(\theta^n, \sigma_{i,1}^2, \sigma_{j,2}^2) \\
&= (p_t^n)^2 N(\theta^n, \sigma_{t,1}^2, \sigma_{t,2}^2) + p_t^n p_f^n N(\theta^n, \sigma_{t,1}^2, \sigma_{f,2}^2) + p_t^n p_w^n N(\theta^n, \sigma_{t,1}^2, \sigma_{w,2}^2) + \\
&\quad p_f^n p_t^n N(\theta^n, \sigma_{f,1}^2, \sigma_{t,2}^2) + (p_f^n)^2 N(\theta^n, \sigma_{f,1}^2, \sigma_{f,2}^2) + p_f^n p_w^n N(\theta^n, \sigma_{f,1}^2, \sigma_{w,2}^2) + \\
&\quad p_w^n p_t^n N(\theta^n, \sigma_{w,1}^2, \sigma_{t,2}^2) + p_w^n p_f^n N(\theta^n, \sigma_{w,1}^2, \sigma_{f,2}^2) + (p_w^n)^2 N(\theta^n, \sigma_{w,1}^2, \sigma_{w,2}^2)
\end{aligned}
$$

Algorithm 2.1 provides a pseudocode representation of the method. We use a uniform prior over the given domain and a diagonal covariance matrix $\hat{\Sigma}$ for our fixed Gaussian. We define N_{step} as the number of time steps and M as the number of parameters. We also define

Algorithm 2.1 Rejection-scaled adaptive proposal (RSAP)

1. Initialize θ^0
2. **for** $n = 1$ **to** N_{step} **do**
3. **if** $n = 1$ **or** $\theta'^{,n-1}$ was accepted **then**
4. Set $\Sigma = \hat{\Sigma}$, $k_w^{n,m} = 0$, $k_t^{n,m} = 0$ for all m
5. **else**
6. **for** $m = 1$ **to** M **do**
7. Choose from {wide, thin, fixed} with probabilities $\{p_w^n, p_t^n, p_f^n\}$
8. **if** wide **then**
9. $k_w^{n,m} = k_w^{n-1,m} + 1$
10. $A_w^m = 1 - (1 - \hat{A}_w)(1 - \exp(-r_w k_w^{n,m}))$
11. $\Sigma_{m,m} = A_w^m \hat{\Sigma}_{m,m}$
12. **else if** thin **then**
13. $k_t^{n,m} = k_t^{n-1,m} + 1$
14. $A_t^m = 1 - (1 - \hat{A}_t)(1 - \exp(-r_t k_t^{n,m}))$
15. $\Sigma_{m,m} = A_{thin}^m \hat{\Sigma}_{m,m}$
16. **else**
17. $\Sigma_{m,m} = \hat{\Sigma}_{m,m}$
18. **end if**
19. **end for**
20. **end if**
21. Generate $\theta'^{,n} \sim N(\theta^{n-1}, \Sigma)$
22. Compute $\alpha(\theta'^{,n}|\theta^{n-1}) = \min\left(1, \frac{\ell(y|\theta'^{,n})p(\theta'^{,n})}{\ell(y|\theta^{n-1})p(\theta^{n-1})}\right)$
23. Set $\theta^n = \theta'^{,n}$ with probability α, else $\theta^n = \theta^{n-1}$
24. **end for**

separate counters k_w^m, k_t^m for every parameter $m = 1, ..., M$. As can clearly be seen, lines 3-20 are the additional adaptive functionality while lines 21-23 are from the standard Metropolis method.

2.3. Ergodicty of RSAP. We will now demonstrate the ergodicity of RSAP by proving that conditions *a* and *b* of Theorem 5 from Roberts and Rosenthal [8] hold for our method. Theorem 5 states

Theorem 2.1 (Theorem 5 [8]). *Consider an adaptive MCMC algorithm, on a state space $\mathcal{X}$, with adaptation index $\mathcal{Y}$, so $\pi(\cdot)$ is stationary for each kernel P_y for $y \in \mathcal{Y}$. Assume that:*

a. (Simultaneous Uniform Ergodicity) *For all $\epsilon > 0$, there $N = N(\epsilon) \in \mathbb{N}$ such that $||P_y^N(x, \cdot) - \pi(\cdot)|| \leq \epsilon$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.*

b. (Diminishing Adaptation) *The limit $\lim_{n\to\infty} D_n = 0$ in probability, where $D_n = \sup_{x\in\mathcal{X}} ||P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_{n+1}}(x, \cdot)||$ is a $\mathcal{G}_{n+1}$-measurable random variable (depending on the random values Γ_n and Γ_{n+1}).*

Then the adaptive algorithm is ergodic.

More plainly, the theorem states that an adaptive method is ergodic if it is (a) *simultaneously uniformly ergodic* and has (b) *diminishing adaptation*. Condition a means that all proposals which could be used in the adaptive method are themselves ergodic if used as proposals for standard Metropolis. Condition b means that the "amount of adaptation" eventually vanishes, i.e., that the sequence of proposals converges (see [8] for more rigorous definitions as well as a proof of Theorem 5). Condition a clearly holds in the case of RSAP since the proposals are all Gaussians, which are ergodic. Therefore, the remainder of this section will be devoted to proving (b).

2.3.1. Proving diminishing adaptation. Suppose there are M parameters which RSAP must estimate and that $\hat{\Sigma}$ is a fixed diagonal covariance matrix for these parameters. First, define the following as the *elemental proposals*

$$E\Big((j_1, i_1), \cdots, (j_M, i_M)\Big) = N(\theta^n, \Sigma)$$

where

$$\Sigma_{m,m} = \begin{cases} \hat{\Sigma}_{m,m} \cdot A_t(i_m) & j_m = 1 \\ \hat{\Sigma}_{m,m} & j_m = 2 \\ \hat{\Sigma}_{m,m} \cdot A_w(i_m) & j_m = 3 \end{cases}$$

with their respective *acceptance probabilities* $a^n_{(j_1,i_1),\cdots,(j_M,i_M)} = \alpha(\theta'|\theta^n)$ where θ' was drawn from the elemental proposal described by the subscripts and $m = 1, \cdots, M$. We let n be the time step, the i_m's be the adaptation degrees, and the j_m's be switches that determine to which mode of adaptation the i_m's are applied (1 = thin, 2 = fixed, 3 = wide). Note that n and i_m are non-negative integers such that $n \geq i_m$ since, in practice, the degree of adaptation cannot exceed the number of time steps which have occured. For the sake of notational simplicity, we will also define the *rejection probabilities* $r^n_{(j_1,i_1),\cdots,(j_M,i_M)} = 1 - a^n_{(j_1,i_1),\cdots,(j_M,i_M)}$.

Next, define the following linear combinations as the *adaptive proposals*

$$q^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M} = \left(\sum_{j_1=1}^{3} p^n_{j_1} \cdots \sum_{j_M=1}^{3} p^n_{j_M} E\Big((j_1, i_1^{j_1}), \cdots, (j_M, i_M^{j_M})\Big) \right) \tag{2.4}$$

with their associated *acceptance probabilities*

$$b^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M} = \left(\sum_{j_1=1}^{3} p^n_{j_1} \cdots \sum_{j_M=1}^{3} p^n_{j_M} a^n_{(j_1,i_1^{j_1}),\cdots,(j_M,i_M^{j_M})} \right)$$

This is the M-parameter generalization of the one- and two-parameter proposals from Section 2.2 with the subscripts t, f, w replaced with $1, 2, 3$, respectively. Let's examine Equation (2.4) more closely before continuing on. Regarding the indices, the integers i^1_m are the degrees of proposal-*thinning* adaptation while the integers i^3_m are the degrees of proposal-*widening* adaptation. The integers i^2_m can be thought of as the degrees of proposal-*fixing* adaptation (clearly a meaningless notion). They are merely auxiliary indices which help to establish conservation of adaptation degrees. Regarding the sums, there is one per parameter and their

indices cover the three adaptation options so that the probability of choosing the elemental proposal $E((j_1, i_1^{j_1}), \cdots, (j_M, i_M^{j_M}))$ is $p_{j_1}^n \cdots p_{j_M}^n$.

Lastly, define Q^n as the *full proposal* at time-step n. Since acceptance and rejection are random phenomena, we cannot know ahead of time which specific elemental or adaptive proposal will be used at time step n. Thus, the full proposal will be a linear combination of the adaptive proposals whose total adaptation degrees are less than or equal to n (since there could only have been up to n rejections and, therefore, n adaptations) with functions of their respective acceptance probabilities.

Let us now examine the structure of the full proposal and how it evolves throughout a run, temporarily restricting ourselves to one dimension for simplicity. Below, Figure 1 provides a graphical representation of the flow of proposals during a run (with the time step superscripts ommitted). One can think of method runs as paths through this graph. The

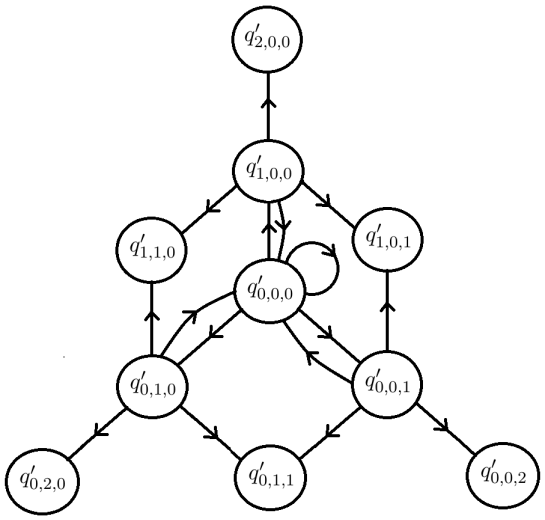

Figure 3. *A graphical depiction of RSAP and the possible transitions that can take place. At any node (adaptive proposal) in the graph, four transitions can occur: increasing one of the three indices (rejection) or returning to the initial state (acceptance). Note that we have omitted the time step superscript since to properly represent the method graphically would require another dimension to represent time. We instead identify all the proposals with the same adaptation degrees and simply use primes to denote the fact that they represent numerous other proposals with equivalent adaptation degrees but larger time step superscripts.*

first state is always $q_{0,0,0}^0$ since no adaptation has yet taken place. Consequently, the first full proposal q^0 is also $q_{0,0,0}^0$. As illustrated by the graph, there are four possible transitions from any given state. If acceptance occurs, the state returns to $q_{0,0,0}^0$ (the fixed proposal with no adaptation). Alternatively, if the candidate is rejected, we will increment the adaptive degree indices corresponding to the elemental proposal which was used. For example, suppose that we chose to use a thin elemental proposal and the sample drawn from it was rejected. Then, we increment the thin proposal index so that the adaptive proposal used in the next time step will be $q_{1,0,0}^1$. We can represent this mathematically with linear combinations of the adaptive

proposals. Equations (2.5) below show the full proposals through q^2.

$$
\begin{aligned}
Q^0 &= q^0_{0,0,0} \\
Q^1 &= b^0_{0,0,0}q^1_{0,0,0} + r^0_{(1,0)}p^0_1 q^1_{1,0,0} + r^0_{(2,0)}p^0_2 q^1_{0,1,0} + r^0_{(3,0)}p^0_3 q^1_{0,0,1} \\
Q^2 &= \left(b^1_{0,0,0}b^0_{0,0,0} + b^1_{1,0,0}r^0_{(1,0)}p^0_1 + b^1_{0,1,0}r^0_{(2,0)}p^0_2 + b^1_{0,0,1}r^0_{(3,0)}p^0_3\right)q^2_{0,0,0} + \\
&\quad r^1_{(1,0)}p^1_1 b^0_{0,0,0}q^2_{1,0,0} + r^1_{(2,0)}p^1_2 b^0_{0,0,0}q^2_{0,1,0} + r^1_{(3,0)}p^1_3 b^0_{0,0,0}q^2_{0,0,1} + \\
&\quad r^1_{(1,1)}p^1_1 r^0_{(1,0)}p^0_1 q^2_{2,0,0} + r^1_{(2,1)}p^1_2 r^0_{(1,0)}p^0_1 q^2_{1,1,0} + r^1_{(3,1)}p^1_3 r^0_{(1,0)}p^0_1 q^2_{1,0,1} + \\
&\quad r^1_{(1,1)}p^1_1 r^0_{(2,0)}p^0_2 q^2_{1,1,0} + r^1_{(2,1)}p^1_2 r^0_{(2,0)}p^0_2 q^2_{2,0,0} + r^1_{(3,1)}p^1_3 r^0_{(2,0)}p^0_2 q^2_{0,1,1} + \\
&\quad r^1_{(1,1)}p^1_1 r^0_{(3,0)}p^0_3 q^2_{1,0,1} + r^1_{(2,1)}p^1_2 r^0_{(3,0)}p^0_3 q^2_{0,1,1} + r^1_{(3,1)}p^1_3 r^0_{(3,0)}p^0_3 q^2_{0,0,2}
\end{aligned}
\tag{2.5}
$$

As can be seen (and was mentioned earlier), Q^n contains all adaptive proposals of total degree up to n, i.e., $i^1_1 + i^2_1 + i^3_1 \leq n$ (for M parameters, this inequality holds for all parameters simultaneously) . We can express Q^n neatly as

$$
Q^n = \sum_{\ell=0}^{n} \left(\sum_{i^1_1+i^2_1+i^3_1=\ell} c^n_{i^1_1,i^2_1,i^3_1} q^n_{i^1_1,i^2_1,i^3_1} \right)
$$

where

$$
\sum_{\ell=0}^{n} \left(\sum_{i^1_1+i^2_1+i^3_1=\ell} c^n_{i^1_1,i^2_1,i^3_1} \right) = 1
$$

defines our normalization (since some proposal must always be chosen) and the inner sums are taken over all possible values of the three indices such that their sum is less than or equal to n. The value $c^n_{i^1_1,i^2_1,i^3_1}$ can be interpreted as the probability that that the adaptive proposal defined by the subscripts will be used at step n.

Now, returning to M parameters, let us generalize the above formulae. The full proposal is written

$$
Q^n = \sum_{\ell=0}^{n} \left(\sum_{i^1_m+i^2_m+i^3_m=\ell} c^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M} q^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M} \right)
$$

with normalization

$$
\sum_{\ell=0}^{n} \left(\sum_{i^1_m+i^2_m+i^3_m=\ell} c^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M} \right) = 1
$$

where $m = 1, \cdots, M$. We can construct several recurrence relations which describe how the values of the coefficients c^n change over time. We get one relation which governs the evolution of the coefficient corresponding to the proposal which has zero total adaptation degree and

another which governs all the rest.

$$c^{n+1}_{0,0,0,\cdots,0,0,0} = \sum_{\ell=0}^{n}\left(\sum_{i^1_m+i^2_m+i^3_m=\ell}\left(b^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M}\right)\left(c^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M}\right)\right)$$

$$c^{n+1}_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M} = \sum_{j_1|i^{j_1}_1\geq 1}\cdots\sum_{j_M|i^{j_M}_M\geq 1}\left(r^n_{(j_1,i^{j_1}_1),\cdots,(j_M,i^{j_M}_M)}\right)\left(p^n_{j_1}\cdots p^n_{j_M}\right)\times \left(c^n_{i^1_1,\cdots,i^{j_1}_1-1,\cdots,i^3_1,\cdots,i^1_M,\cdots,i^{j_M}_M-1,\cdots,i^3_M}\right) \tag{2.6}$$

The first relation is fairly basic, as it simply was derived from the fact that the adaptation-less proposal will be used at any step which immediately follows an acceptance. Consequently, it finds the total, weighted acceptance probability given all possible proposals which could be used at that step. This is equal to the probability that the adaptation-less proposal will be chosen at the following step. The second relation is more complicated. Conceptually, it finds the total, weighted rejection probability from all proposals which could possibly precede the proposal under consideration. More specifically, it does this by summing the weighted rejection probabilities over all valid permutations of subtracting one from the three types adaptation degrees for all M parameters (of which there are no more than 3^M permutations). A permutation is considered valid when subtracting one does not give a negative adaptation degree (thus the $i^{j_m}_m \geq 1$ under the sums).

Now, to prove diminishing adaptation, we will show that the values of $c^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M}$ vanish as $n \to \infty$ for any proposals that have at least one degree of either thinning or widening adaptation in at least one parameter. We will use induction on the thinning and widening adaptation degrees to prove this.

Let us first prove the base case, where we consider a proposal such that each parameter has only one adaptation degree assigned and there exists at least one parameter such that its mode of adaptation is either thinning or widening. Without loss of generality, we will say that this is the first parameter and that thinning was chosen (the widening case is identical). Let $c^n_{1,0,0,\cdots,i^1_M,i^2_M,i^3_M}$ be the coefficient of this proposal. Then, using the recurrence relation, we get

$$c^{n+1}_{1,0,0,\cdots,i^1_M,i^2_M,i^3_M} = \left(r^n_{(1,0),\cdots,(j_M,i^{j_M}_M)}\right)\left(p^n_1\cdots p^n_{j_M}\right)\left(c^n_{0,0,0,\cdots,0,0,0}\right)$$

Now, all the terms in this equation are bounded between 0 and 1 and $p^n_1 \to 0$ as $n \to \infty$. Therefore, $c^{n+1}_{1,0,0,\cdots,i^1_M,i^2_M,i^3_M} \to 0$, as well, and we have that all coefficients with at least one parameter having a single adaptive degree assigned to either thinning or widening vanish.

We will now prove the induction step, which states that if any coefficient with at least one parameter having at least one thinning or widening adaptive degree vanishes, then the coefficients for all proposals that it precedes by one time step (in the sense of the discussion of Equation (2.6)) do as well. Consider $c^n_{i^1_1,i^2_1,i^3_1,\cdots,i^1_M,i^2_M,i^3_M}$, which has at least one parameter with at least one thinning or widening adaptive degree. Equation (2.6) applied to this coefficient contains the coefficients for all possible preceding states. Now, since there is at least one degree of thinning or widening adaptation in the coefficient under consideration, every term in the

sum from Equation (2.6) can be classified in one of two ways: either 1) that adaptation degree was obtained from the immediately preceding step or 2) from an earlier step. If it is from the immediately preceding step, we can factor out the appropriate probability term as we did for the base case, causing that term to vanish. If it was from an earlier step, then the proposal associated with that term already has at least one parameter with at least one thinning or widening adaptive degree, and thus—according to the induction hypothesis—vanishes. Having shown that every term in the sum vanishes, we have shown that all proposal coefficients with at least one parameter with at least one thinning or widening adaptive degree also vanish. Thus, the only non-vanishing coefficients are those for purely fixing proposals, which are all identical to the adaptation-less proposal. Therefore, the full proposal $Q^n \to q^n_{0,0,0,\cdots,0,0,0}$ as $n \to \infty$, thus proving part (b) of Theorem 2.1, i.e., diminishing adaptation. Therefore, having proven both (a) and (b), we have also proven the ergodicity of RSAP.

It is readily apparent that we took advantage of the ergodicity of Metropolis in order to prove the ergodicity of RSAP. In fact, the convergence of the mixing probabilities to $p^n_f \to 1$ essentially forces the method into Metropolis over time. We suspect, however, that RSAP might be ergodic without this condition on the mixing probabilities. Proving this would require showing that Equation (2.6) implies that the c^n's converge in distribution. They certainly do not directly converge in value since they depend on the highly-stochastic acceptance/rejeciton probabilities; however, an argument can likely be made that this implies their convergence in distribution. We will leave this for a future paper since the current version of RSAP is sufficient.

3. Computational testing.

Having proven the ergodicity of our method, we now move on to demonstrate its practical utility through the presentation of numerous computational results. These cover multiple different aspects of the method's performance, including flexibility, mixing efficiency, and speed of convergence. We also present results from a new optimization-style test we developed in which we find the cumulative probability distribution for convergence of the chain to the global minimum of the test function versus a range of fixed proposal widths. This technique provides easy-to-interpret results which are numerically *and* visually simple to compare. This test will be applied to RSAP, Metropolis, and modified Adaptive Metropolis from Section 1.3.

3.1. Discussion of RSAP and modified AM parameter values used in testing.

When performing numerical experiments on our method, we used standard optimization test functions, each with a global minimum of $y = f(\theta^*) = 0$ (the bi-modal function has two adjacent minima). As such, the likelihood function can be expressed as

$$\begin{aligned}\ell(y|\theta) &\propto \exp\left(-\frac{1}{2}\frac{(y-f(\theta))^2}{\Delta^2}\right)\\ &\propto \exp\left(-\frac{1}{2}\frac{f(\theta)^2}{\Delta^2}\right)\end{aligned}$$

The value of Δ is related to the scale of the error. It also affects the ease of acceptance, and thus the relative peakedness of the resulting sample-binned approximation of the posterior. Not knowing its proper value, we include it as a parameter for RSAP to estimate, though we

set its proposal width quite low so that it does not change too rapidly, dramatically altering the acceptance rate (we further discuss this parameter and its effects on performance in Section 3.3).

There is also the matter of setting the adaptation parameters to their proper values. Clearly, the optimal values of these parameters are problem-dependent; so we attempted to find a set of reasonable values which improved the performance gain over Metropolis for as many test cases as possible. We thus arrived at maximum scaling magnitudes of $\hat{A}_t = 0.1, \hat{A}_w = 10$, respectively. Being able to both increase and decrease the proposal width by an order of magnitude ensures that the true optimal width is likely contained within the interval $[\hat{A}_t\sigma_f, \hat{A}_w\sigma_f]$. Next, concerning the scaling rates, we found that their values were less significant than the magnitudes in determining performance gain. Their values were set to $r_t = r_w = 0.3$, since this puts the scaling to within approximately 5% of its maximum effect within 10 rejection steps ($e^{-0.3\cdot 10} \approx 0.049$). Smaller rate values would require more consecutive rejections to reach the same amount of scaling, thus leaving the more extreme scaling values unused. Lastly, the interval widths n_1, n_2 are likely the most problem-dependent parameters since they determine how long RSAP is allowed to be adaptive. As discussed earlier, n_1 should be large w.r.t. N_{step} while $n_2 \leq N_{step} - n_1$ has more freedom. Because of their problem dependence, we will vary them throughout the tests.

Concerning modified AM, we ran the method for N_{burn} steps using the same fixed proposal widths as the other two methods. Once these steps have passed, we begin using the modified AM kernel, while ignoring a specified fraction of the initial steps (specifically, the first $N_{burn}/2$ steps) in all of the covariance calculations. This is so that the covariance matrix calculations aren't as biased by the choice of initial state. Also, we set $\beta = 0.5$ for all of the tests.

3.2. Chain mixing. To begin the discussion of our results, we will provide trace plots of one-parameter chains created by RSAP for the test functions under consideration, as well as plots of the functions used and sample-binned approximations of the posterior distribution. We will start with the 1D Gaussian defined by

$$f(\theta) = a\Big(1 - \exp(-\frac{1}{2}\frac{\theta^2}{b^2})\Big)$$

where $a = 1, b = 0.3, \theta \in [-1, 1]$. The initial state values were set to $\theta^0 = 0.8, \Delta^0 = 0.08$ and the fixed proposal widths were set to $\sigma_\theta = 0.1, \sigma_\Delta = 0.0008$. As discussed above, the parameters for RSAP were set to $\hat{A}_t = 0.1, \hat{A}_w = 10, r_t = 0.3, r_w = 0.3, n_1 = 2000, n_2 = 1000$. We use the Gaussian as a control example since it is such a simple function and does not require adaptation.

Next, we have an Ackley function in 1D defined by

$$f(\theta) = a\Big(1 - \exp(-0.2\sqrt{\theta^2/M})\Big) + b\Big(e - \exp(\cos(2\pi\theta)/M)\Big)$$

where $a = 20, b = 4, M = 1, \theta \in [-10, 10]$. The initial state values were set to $\theta_0 = 8, \Delta_0 = 3.5$ and the fixed proposal widths were set to $\sigma_\theta = 1.0, \sigma_\Delta = 0.01$. The RSAP parameters again were set to $\hat{A}_t = 0.1, \hat{A}_w = 10, r_t = 0.3, r_w = 0.3, n_1 = 2000, n_2 = 1000$. It is very easy to see from the trace and posterior plots that RSAP efficiently samples the modes near the center.

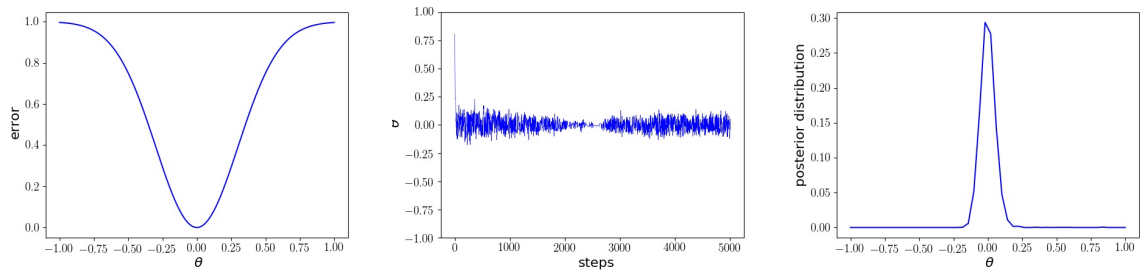

Figure 4. *Plots of the Gaussian function (left), chain trace (center), and binned posterior estimate (right) resulting from a run with 5,000 steps. For the posterior plot on the right, we used 50 bins, giving a bin width of 0.04.*

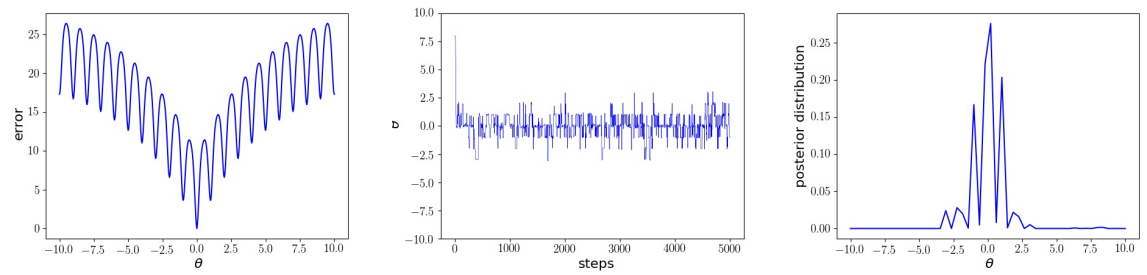

Figure 5. *Plots of the Ackley function (left), chain trace (center), and binned posterior estimate (right) resulting from a run with 5,000 steps. For the posterior plot on the right, we used 50 bins, giving a bin width of 0.4.*

Lastly, we have a rather illustrative example of the superiority of RSAP to standard Metropolis in the form of a bi-modal function defined by

$$f(\theta) = a\Big(1 - \exp(-\frac{1}{2}\Big(\frac{(\theta - c)^2}{b^2}\Big)^{d/2})\Big) + a\Big(1 - \exp(-\frac{1}{2}\Big(\frac{(\theta + c)^2}{b^2}\Big)^{d/2})\Big)$$

where $a = 0.5, b = 0.15, c = 0.333, d = 8, \theta \in [-1, 1]$. The initial state values were set to $\theta^0 = 0.8, \Delta^0 = 0.08$ and the fixed proposal widths were set to $\sigma_\theta = 0.1, \sigma_\Delta = 0.0008$. As discussed above, the parameters for RSAP were set to $\hat{A}_t = 0.1, \hat{A}_w = 10, r_t = 0.3, r_w = 0.3, n_1 = 2000, n_2 = 1000$. As can be seen, there is efficient mixing between the two modes in the region up until $n = n_1$; however, once standard Metropolis begins to overtake the adaptive effects, the chain becomes temporarily "stuck" in whichever mode it was currently in. One can simply increase n_1 to a suitably large value to maintain the level of mixing efficiency for a longer period of time.

Obviously, this test can be easily skewwed by choosing method parameters which work well for RSAP but poor for Metropolis or AM. The tests in the next section provide a more fair comparison, letting us examine the optimal performance of each method.

3.3. Convergence to global minimum. We now move on to present results from our optimization-inspired experiments. These use statistics from large ensembles of chains to provide a systematic way of comparing the short-term (i.e., burn-in) performance of Metropolis,

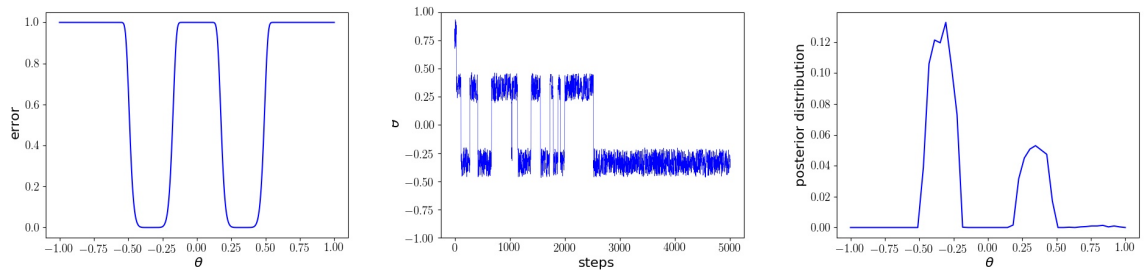

Figure 6. *Plots of the bi-modal function (left), chain trace (center), and binned posterior estimate (right) resulting from a run with 5,000 steps. For the posterior plot on the right, we used 50 bins, giving a bin width of 0.04.*

Adaptive Metropolis, and RSAP over different proposal widths. This is useful because it allows us to compare both methods in their optimal context, since sweeping through a large range of proposal widths is guaranteed to contain a near-optimal value. As we shall soon see, RSAP outperforms both Metropolis and Adaptive Metropolis for most proposal widths, particularly in higher dimensionalities.

When performing this test (see Algorithm 3.1), we used M-dimensional Ackley functions

$$f(\theta) = 20\left(1 - \exp\left(-0.2\left(\frac{1}{M}\sum_{m=1}^{M}\theta_m^2\right)^{0.5}\right)\right) + \left(e - \exp\left(\frac{1}{M}\sum_{m=1}^{M}\cos(2\pi\theta_m)\right)\right)$$

over a domain defined by the M-dimensional cube $[-L, L]^M$. We sweep through N_{width} proposal widths at equally-spaced intervals, i.e., $\sigma^i = i\sigma_{max}/(N_{width} - 1)$ where $0 \leq i \leq N_{width} - 1$. Obviously, since RSAP adapts the proposal width, we will be sweeping through its *fixed* proposal width, letting these values also equal the widths for Metropolis. We say "widths" rather than "covariance matrices" since—for the Ackley function—each parameter axis is scaled equally, thus causing a single width to work for each. For every width value, we run N_{chain}, each for N_{step}, with initial states that are uniformly-distributed across the cubical domain. The cumulative probability of convergence to the global minimum of the Ackley function versus the time-step is then calculated for each width by determining at which time-step each chain's Ackley evaluation is below the threshold ε.

Let us note two more considerations before presenting the results. Since we are merely concerned with the short-term behavior of RSAP, the values n_1, n_2 play no role, thus placing us in the adaptive regime the entire time. Also, concerning the value of Δ, lower values are superior for fast initial convergence but not necessarily ideal for accurate sampling of the posterior in the long-term. RSAP, in particular, prefers a low value of Δ, since doing so causes more rejections to occur, letting RSAP adapt more frequently. Thus, we will restrict ourselves to only small values of Δ.

Now, let us examine the results of our numerical tests. Figure 7 and Table 1 contain the results of a comparison between Metropolis, Adaptive Metropolis, and RSAP on an Ackley function in 3, 10, and 20 dimensions. For the 3D test (top row), there are several features which must be discussed. Observing Figure 7, we notice that both Metropolis and RSAP have

Algorithm 3.1 Test for convergence to global max

1. Define the M-dimensional parameter vector $\theta^{n,i,j}$ as the n-th time-step of the j-th chain of the i-th proposal width
2. **for** $i = 1$ **to** N_{width} **do**
3. **for** $j = 1$ **to** N_{chain} **do**
4. Initialize chain: $\theta^{0,i,j} \sim U([-L, L]^M)$ (note that $\theta^{0,i,j}$ is vector-valued)
5. Calculate the chain $\{\theta^{0,i,j}, \cdots, \theta^{n,i,j}\}$ for $n = 1, ..., N_{step}$ steps using either Metropolis or RSAP with initial state $\theta^{0,i,j}$ and fixed proposal covariance matrix $\Sigma^i = \sigma^i I$
6. **end for**
7. Initialize the cumulative distribution: $c_0^i = 0$
8. **for** $j = 1$ **to** N_{chain} **do**
9. Initialize convergence check: isConv = false
10. **for** $n = 1$ **to** N_{step} **do**
11. **if not** isConv **and** $f(\theta^{n,i,j}) \leq \varepsilon$ **then**
12. $c_n^i = c_{n-1}^i + 1$
13. isConv = true
14. **else**
15. $c_n^i = c_{n-1}^i$
16. **end if**
17. **end for**
18. **end for**
19. **end for**
20. Normalize c by dividing all entries by N_{chain}

	Max fraction converged at given run-time completion %											
	3D				10D				20D			
run-time %	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%
Metro.	0.09	0.33	0.60	0.75	0.01	0.03	0.10	0.15	0.00	0.00	0.00	0.00
AM	0.06	0.25	0.73	0.78	0.01	0.16	0.24	0.30	0.00	0.00	0.00	0.00
RSAP	0.19	0.73	0.96	0.99	0.20	0.84	0.99	1.00	0.01	0.33	0.84	0.99

Table 1

Numeric convergence data from Figure 7. The values are the maximum fraction of converging runs across 4 different time-slices of Figure 7 (i.e., the maximum is taken over the range of proposal widths used). The data is grouped by the method used (rows), as well as dimensionality of the problem (columns) and run-time completion percentage (columns). Values were rounded to the nearest hundredth.

relatively similarly shaped convergence profiles, with the fastest rate of convergence happening near some particular optimal value of the proposal width and a tapering off as the width is increased or decreased. The main differences between the two are 1) that RSAP's optimal width is slightly less than that of Metropolis and, more obviously, 2) that RSAP's convergence rate is significantly higher than Metropolis's for all width values. The is confirmed by Table 1, where we see that RSAP had a near perfect convergence rate within 500 steps while Metropolis only had roughly 75% convergence. We also notice that many of the RSAP runs converged much earlier in time than those for Metropolis.

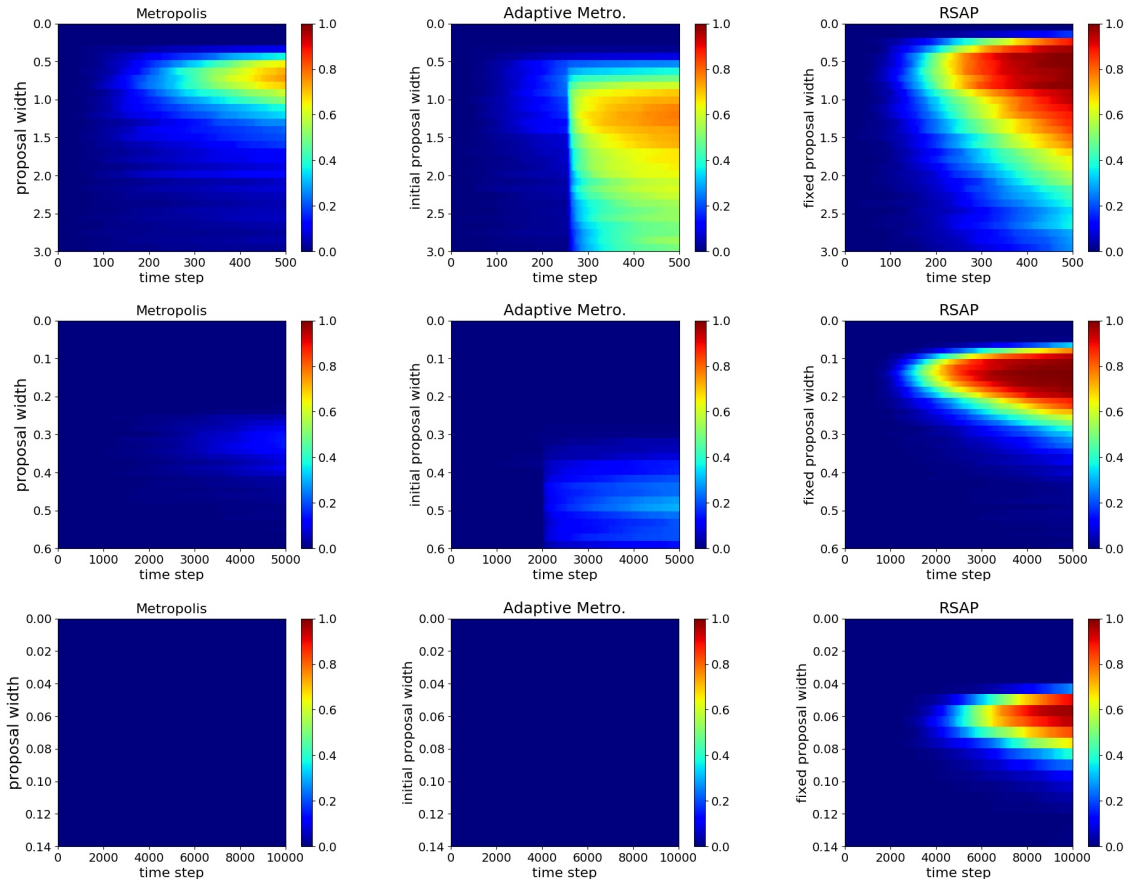

Figure 7. *Plots of the cumulative convergence of Metropolis (left), Adaptive Metropolis (center), and RSAP (right) performed on a 3D (top), 10D (center), 20D (bottom) Ackley function. For 3D, we used* $N_{burn} = 250$, $\Delta = 0.01$, *5000 time-steps per chain, 500 chains per width value, and threshold of 1.0 over the domain* $[-15.0, 15.0]^3$. *For 10D, we used* $N_{burn} = 2000$, $\Delta = 0.01$, *5000 time-steps per chain, 500 chains per width value, and threshold of 1.0 over the domain* $[-15.0, 15.0]^{10}$. *For 20D, we used* $N_{burn} = 5000$, $\Delta = 0.001$, *10000 time-steps per chain, 500 chains per width value, and threshold of 1.0 over the domain* $[-10.0, 10.0]^{20}$.

On the other hand, the convergence profile for AM is quite different from both Metropolis and RSAP, with its transition into the adaptive regime being very clearly visible at roughly 250 steps. Since AM's proposal width (rather, covariance matrix) should converge to the same value over a large range of initial values, we see similar performance across a wide range of initial widths. Also, though AM's convergence does have an optimal width value, it is much more forgiving than the other two methods when using a suboptimal width. Comparing with Metropolis, Table 1 shows that AM's optimal widths aren't significantly better than standard Metropolis; rather—as previously stated—AM's true benefit lies in its incredibly consistent convergence across a range of different widths. Also, AM outperforms RSAP for larger width values; however, RSAP is still able to outperform both methods when given widths which are smaller than ideal for Metropolis or AM.

Considering just these small width values, it is reasonable for Metropolis and AM to

stagnate on the Ackley function if given too small a width. Each trough in the Ackley function has a width of 1.0 between adjacent modes. This is due to the $\cos(2\pi\theta_m)$ term. Now, for a Gaussian distribution, 99.7% of samples drawn will be within 3σ of the mean. Therefore, letting $1.0 = 3\sigma$, we see that a width of 0.333 will only draw samples outside of a particular trough 0.3% of the time (assuming that the mean is approximately at the center of the trough, which is reasonable since it will be attracted to that location). If Metropolis were to use such a small width and become trapped in one of Ackley's modes, it would likely take a very large number of steps to escape. Similarly for AM, if its estimation of the optimal width ever became this small, drawing further samples from within the trough would only serve to *decrease* the width estimate even further (AM would have a better chance of escaping such a trough in a low number of dimensions since the coefficient $2.38^2/d > 1$, thus causing the proposal width value to be larger than 0.333). Looking back to the top row of Figure 7 (3D), we see no convergence in Metropolis or AM for width values at or below 0.333, with AM actually performing worse than standard Metropolis. On the other hand, due to RSAP's ability to both grow and shrink the width, it still converges well within this region.

Moving on, we have a comparison over a 10D Ackley function (center row). Note the scale of the axes has changed dramatically from the 3D case since a 10D space requires more time-steps to explore and the optimal width decreases [10]. In spite of the change of scale, the shape of the three methods' convergence profiles remains similar to their lower-dimensional counterparts, with all three methods having a clear optimal width and AM have more consistent performance across a range of widths, as well as roughly twice the convergence fraction of Metropolis. However, now that we're considering a 10D problem, the superiority of RSAP becomes quite evident, with 100% convergence of runs within 10000 steps while on Metropolis and AM only achieve 15% and 30%, respectively. Also of note is that RSAP's optimal width is now significantly smaller that those of Metropolis or AM.

Finally, we have a comparison over a 20D Ackley function (bottom row). Note again that the axes' scales have changed again). For this test, we shrunk the domain from $[-15.0, 15.0]^{20}$ to $[-10.0, 10.0]^{20}$ and decreased Δ_0 by a factor of 10. Though it is difficult to tell from the image, we see from the table that there was zero convergence from any of the Metropolis *or* AM runs within 20000 steps. RSAP, on the other hand, still performed quite well, though the range of quickly-converging widths has been greatly compressed. It is clear that the convergence rate of Metropolis and AM in high-dimensional spaces pales in comparison to that of RSAP.

4. Conclusions.

From both our analytical and numerical results, it is clear that RSAP deserves a well-earned a place in the vast marketplace of adaptive MCMC methods. Though it performs at an average rate in low-dimensions, its true power lies in its fast convergence in high-dimensional, multimodal spaces, performing vastly superior to Metropolis and AM in this context.

There are several modifications and applications of RSAP which we would like to investigate in future work. Certainly, different adaptation rules—such as decrementing instead of reseting upon acceptance, scaled incrementing/decrementing based on the acceptance rate, or some type of stochastic incrementing—would be an interesting topic. We also have several ideas on how to combine RSAP and AM in such a way that the resulting method will have

the fast convergence of RSAP along with the consistent convergence across a wide range of widths of AM. Lastly, RSAP is an excellent tool for fitting systems of differential equations to some given data. As such, we have an application for RSAP which involves the parameter estimation/best-fit modeling of galaxy mergers. This has been on ongoing project [12, 13] for a number of years now and the application of RSAP will allow us to automate portions of the workflow which currently take a large number of man-hours to complete.

Acknowledgments. We would like to acknowledge...

REFERENCES

[1] A. Gelman, J ,Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*, 3rd ed., CRC Press, 2014.
[2] H. Haario, E. Saksman, and J. Tamminen, *Adaptive proposal distribution for random walk Metropolis algorithm*, Comput. Stat., 14 (1999), pp. 375-395.
[3] H. Harrio, E. Saksman, J. Tamminen, *An Adaptive Metropolis Algorithm*, Bernoulli, 7(2), 2001, pp. 223-242.
[4] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97-109.
[5] A. Holincheck, J. Wallin, K. Borne, L. Fortson, C. Lintott, A M. Smith, S. Bamford, W. Keel, and M. Parrish, *Galaxy Zoo: Mergers - Dynamical Models of Interacting Galaxies*, MNRAS, 459(1) (2016), pp. 720-745.
[6] L. Jing and P. Vadakkepat, *Interacting MCMC particle filter for tracking maneuvering target*, Digital Signal Processing, 20(2) (2010), pp. 561-574.
[7] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *"Equation of state calculations by fast computing machines*, Journal of Chemical Physics, 21 (1953), pp. 1087-1092.
[8] G. O. Roberts, A. Gelman, and W. R. Gilks, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, Ann. Appl. Probab., 7 (1997), pp. 110-120.
[9] G. O. Roberts and J. S. Rosenthal, *Examples of Adaptive MCMC*, Journal of Computational and Graphical Statistics, 18:2 (2012), pp. 349-367.
[10] G. O. Roberts and J. S. Rosenthal. *Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms*, Journal of Applied Probability, 44(2) (2007), pp. 458-475
[11] G. O. Roberts, and J. S. Rosenthal, *Examples of adaptive MCMC*, Journal of Computational and Graphical Statistics, 18(2) (2009), pp. 349-367.
[12] C. Sherlock, P. Fearnhead, G. O. Roberts, *The Random Walk Metropolis: Linking Theory and Practice Through a Case Study*, Statist. Sci., 25 (2010), no. 2, pp. 172-190.
[13] J. Wallin, A. Holincheck, and A. Harvey, *JSPAM: A restricted three-body code for simulating interacting galaxies.*, Astronomy and Computing, 16 (2016), pp. 26-33.