A Correctness Analysis of the Stochastic Gradient Langevin Descent Algorithm

Raiden Hasegawa

May 21, 2013

1 Introduction

In this short paper I investigate whether the Stochastic Gradient Langevin Descent (SGLD) algorithm presented in [7] for Bayesian posterior sampling is exact or not. A straightforward error analysis suggests that the method will be biased and the number of expected rejections will be on the order of $O(N/n)$ even when the step size is annealed to zero, where N is the size of the dataset and n is the size of the batch. Several experiments with the same toy model presented in the original paper corroborate this and show that the Monte Carlo estimates produced by SGLD are very sensitive to the batch size and the choice of the annealing schedule. The experiments do suggest that SGLD estimates may improve as as the ratio N/n decreases.

2 Algorithm and Analysis

2.1 *Langevin Monte Carlo (LMC)*

Consider a Bayesian model with parameters θ, data $X = \{X_i\}_{i=1}^N$ and a posterior that can be expressed as

$$\mathcal{P}(\theta|X) = \mathrm{P}(\theta) \prod_{i=1}^{N} \mathrm{P}(X_i|\theta) \; . \tag{1}$$

This implies that the likelihood function $\mathcal{L}(X|\theta)$ is separable over the data which is a reasonable assumption. The Monte Carlo proposal using Langevin dynamics can be expressed as

$$\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \left\{ \nabla \log\{\mathrm{P}(\theta)\} + \sum_{i=1}^{N} \nabla \log\{\mathrm{P}(X_i|\theta)\} \right\} + \eta \, , \quad \eta \sim \mathcal{N}(0, \epsilon) \, . \tag{2}$$

Since we end up with a discretization error of $\epsilon^{3/2}$ (which I prove below), a Metropolis Hastings (MH) rejection step needs to be introduced to ensure detailed balance.

2.2 *Stochastic Gradient Langevin Descent (SGLD)*

In a sampling setting where our likelihood involves the evaluation over a massive number of data points, (2) becomes infeasible. At every step we need to evaluate the full likelihood (or gradient thereof) to make the step and then again to decide whether or not to reject. Given Monte Carlo solutions converge at a rate of $O(1/\sqrt{T})$ [3], (2) is only useful for relatively small problems. [7] proposes a potential remedy to this "curse of big data" problem by approximating (2) using ideas from Stochastic Gradient Descent (SGD). They make a few claims that the noise from approximating the gradient in (2) with a random batch of the full data is eventually dominated by the injected noise term η when ϵ follows a polynomial annealing schedule. Thus , SGLD should eventually sample from the correct distribution. Critical to their claim is that the discretization error will go to zero as ϵ is

annealed to zero, thus eliminating the need for a costly MH step. The algorithm can be summarized by the following proposal step

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left\{\nabla \log\{\mathrm{P}(\theta)\} + \frac{N}{n}\sum_{i=1}^{n} \nabla \log\{\mathrm{P}(X_i|\theta)\}\right\} + \eta\,, \quad \eta_t \sim \mathcal{N}(0, \epsilon_t)\,. \tag{3}$$

A batch size of n random data is used to approximate the likelihood gradient. The step size, ϵ_t follows a polynomial annealing schedule such that the SGLD should end up near a mode.

2.2.1 *Concerns*

At first look, this approximation looks promising given the success of SGD in maximization problems [2]. The gist of their claim is that SGLD takes the best of both worlds: efficient maximization using SGD while generating a sample from the Bayesian posterior. But, a closer analysis reveals some potential problems. First, the fact that the algorithm avoids the costly MH step because the discretization error of the SGLD step goes to zero as ϵ is annealed is not a free lunch. The cost for avoiding the MH step is extremely long autocorrelation times since we end up with very small steps. I observe this in the results of my experiments. Secondly, the probability of rejection may go to zero with ϵ but it is not the case that the expected number of rejections over a sampling run will asymptotically approach zero as well. In fact, I show below that one should expect a non-trivial number of rejections over a sampling run, introducing bias into our Monte Carlo estimates.

2.3 *Error Analysis*

I will first show that as $\epsilon \to 0$ the expected number of rejections goes to zero when the full Posterior is evaluated at each Langevin proposal. Then, we will show why this does not hold for SGLD and that one should expect this method to be biased. Consider the Hamiltonian where the negative log-posterior is the potential energy function:

$$H(\theta, \rho) = -\log\{\mathcal{P}(\theta|X)\} + \frac{1}{2}\rho^2\,. \tag{4}$$

The join probability function can be expressed as $\frac{1}{Z}\exp\{-H(\theta,\rho)\}$. Recall that the time dynamics of θ and ρ can be related to the hamiltonian defined in (4) by the following Hamiltonian equations:

$$\frac{d\theta}{dt} = \rho \quad , \quad \frac{d\rho}{dt} = \frac{\partial \log\{\mathcal{P}\}}{\partial \theta} \tag{5}$$

For simplicity, we will work with the case where θ and ρ are of dimension 1. The uncorrected Langevin proposal has the following steps

1. draw ρ from $\mathcal{N}(0, 1)$.
2. Make the following Langevin update (equivalent to a single Leapfrog update for a Hamiltonian Monte Carlo scheme [4]:

$$\theta^* = \theta + \frac{\epsilon}{2}\frac{\partial \log\{\mathcal{P}\}}{\partial \theta}(\theta) + \sqrt{\epsilon}\rho \tag{6}$$

$$\rho^* = \rho + \frac{\sqrt{\epsilon}}{2}\frac{\partial \log\{\mathcal{P}\}}{\partial \theta}(\theta) + \frac{\sqrt{\epsilon}}{2}\frac{\partial \log\{\mathcal{P}\}}{\partial \theta}(\theta^*) \tag{7}$$

$$\tag{8}$$

When we add a MH step to correct for the discretization error in moving along the Hamiltonian in step (2), both steps preserves the invariant distribution. Step (2) can be rewritten as the three steps

of the symplectic Leapfrog integrator which is reversible and preserves volume [4].Now see that the taylor expansion of $\theta(t+\sqrt{\epsilon})$ and $\rho(t+\sqrt{\epsilon})$ at t can be written as

$$\theta(t+\sqrt{\epsilon}) = \theta(t) + \sqrt{\epsilon}\frac{d\theta}{dt}(t) + \frac{\epsilon}{2}\frac{d^2\theta}{dt^2}(t) + O(\epsilon^{3/2}) \tag{9}$$

$$\rho(t+\sqrt{\epsilon}) = \rho(t) + \sqrt{\epsilon}\frac{d\rho}{dt}(t) + \frac{\epsilon}{2}\frac{d^2\rho}{dt^2}(t) + O(\epsilon^{3/2}) \tag{10}$$

Using the Hamiltonian differential equations in (5) and noting that

$$\frac{d}{dt}\frac{d\theta}{dt} = \frac{d\rho}{dt} = \frac{\partial \log\{\mathcal{P}\}}{\partial\theta}$$

we see that the discretization error for θ is

$$\theta^* - \theta(t+\sqrt{\epsilon}) = O(\epsilon^{3/2})\,. \tag{11}$$

The analysis for ρ is a little trickier but the end result is the same in terms of the error. Consider the taylor expansion of $\frac{\partial \log\{\mathcal{P}\}}{\partial\theta}$

$$\begin{aligned}\frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(t+\sqrt{\epsilon}) &= \frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(t) + \sqrt{\epsilon}\frac{\partial}{\partial t}\frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(t) + O(\epsilon)\\ \implies \frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(\theta^*) &= \frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(\theta) + \sqrt{\epsilon}\frac{\partial}{\partial t}\frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(\theta) + O(\epsilon)\end{aligned} \tag{12}$$

The second line holds since the first of our Langevin equations for θ defines the following map $(t, t+\sqrt{\epsilon}) \mapsto (\theta, \theta^*)$. Inserting (12) into (7) gives us

$$\begin{aligned}\rho^* &= \rho + \sqrt{\epsilon}\frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(\theta) + \frac{\epsilon}{2}\frac{\partial}{\partial t}\frac{\partial \log\{\mathcal{P}\}}{\partial\theta}(\theta) + O(\epsilon^{3/2})\\ &= \rho + \sqrt{\epsilon}\frac{\partial\rho}{\partial t}(t) + \frac{\epsilon}{2}\frac{\partial^2\rho}{\partial t^2}(t) + O(\epsilon^{3/2})\,.\end{aligned} \tag{13}$$

Thus, we see that the discretization error for ρ is also of the form

$$\rho^* - \rho(t+\sqrt{\epsilon}) = O(\epsilon^{3/2})\,. \tag{14}$$

When $H(\theta,\rho)$ is well behaved with respect to θ and ρ (e.g. Lipschitz) then our discretization error for $H(\theta,\rho)$ should also be $O(\epsilon^{3/2})$ leading to rejection probabilities of the same order of magnitude. Since LMC exhibits random walk behavior, the number of steps to traverse some given distance will be $O(1/\epsilon)$ [5]. Thus, the expected number of rejections should be $O(\sqrt{\epsilon})$ which goes to zero as ϵ goes to zero.

2.3.1 *Bias in SGLD*

The above analysis suggests we should get an unbiased estimate using LMC without an MH step when our ϵ is small. Although not important to the current point I am trying to make, this observation isn't terribly useful since the growth in autocorrelation time may offset any computational gains from avoiding a rejection step. Back to the point I am trying to make, SGLD as defined in (3) will have discretization error dominated by the approximated gradient, which will be of order $O(\epsilon(N/n))$. This is troublesome in that, the expected number of rejections as $\epsilon \to 0$ is no longer zero. It will be nontrivial and of order $O(N/n)$. This suggests that even if we anneal our step size to zero, SGLD will be asymptotically biased. In the following section I run some experiments that seems to validate this claim.

3 Experiment

In this section I will briefly describe the tests of correctness for SGLD. In summary, I compute the "true" probability of three different regions of a toy posterior and compare the probability derived from the SGLD sampler with varying parameters. As a baseline, I also generate draws using a corrected LMC sampler which in theory should be exact.

3.1 *Toy Model*

The toy model used in [7] is a posterior defined by Gaussian priors and a likelihood that is a mixture of two Gaussians. The model is parameterized in such a way that the two parameters we are sampling are very negatively correlated and the bivariate posterior tends to be bimodal, depending of course, on the data. The model is as follows

$$\theta_1 \sim \mathcal{N}(0, \sigma_1^2); \quad \theta_2 \sim \mathcal{N}(0, \sigma_2^2)$$
$$x_i \sim \frac{1}{2}\mathcal{N}(\theta_1, \sigma_x^2) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, \sigma_x^2)$$

where $\sigma_1^2 = 10$, $\sigma_2^2 = 1$, and $\sigma_x^2 = 2$. The data are generated from the likelihood model where $\theta_1 = 0$ and $\theta_2 = 1$. As mentioned in ([7]. This construction should lead to a bimodal distribution as the number of data N increases, with modes at $(\theta_1, \theta_2) = (0, 1)$ and $(\theta_1, \theta_2) = (1, -1)$ since the likelihood function is invariant under these two cases.

3.2 *Test Statistic*

Three regions, $\{R_i : i \in [1, 3]\}$, of the posterior distribution for θ_1 and θ_2 are chosen to construct our test statistics used to evaluate the exactness of the SGLD algorithm. Using a trapezoidal integration scheme with a 1000×1000 grid I compute $\Pr[R_i]$ for each i. Since the normalizing constant Z is analytically intractable, we approximate it by using the same integration scheme over a large area that should contain nearly all the probability mass. Since the trapezoidal rule has global error of order 2, our grid over $[-10, 10] \times [-10, 10]$ should result in an error with order of magnitude 10^{-4} which should be plenty of precision for our tests. We should expect errors on the order of 10^{-3} from our correct sampling schemes when generating 10^6 draws, and even larger when considering the autocorrelation time of the sequence of draws. The results suggest errors that are closer to 10^{-2} for the correct sampler. Our sampling estimates can be written as

$$\widehat{\Pr[R_i]} = \frac{\sum_{k=1}^{n} \mathbb{1}_{R_i}(X_k)}{\sum_{k=1}^{n} \mathbb{1}_{[-10,10]^2}(X_k)} \tag{15}$$

where the denominator corrects for the fact that our true value for $\Pr[R_i]$ is in fact a conditional probability. Since our $\sum_{k=1}^{n} \mathbb{1}_{R_i}$ follows a binomial distribution one can write the standard errors of $\widehat{\Pr[R_i]}$ as

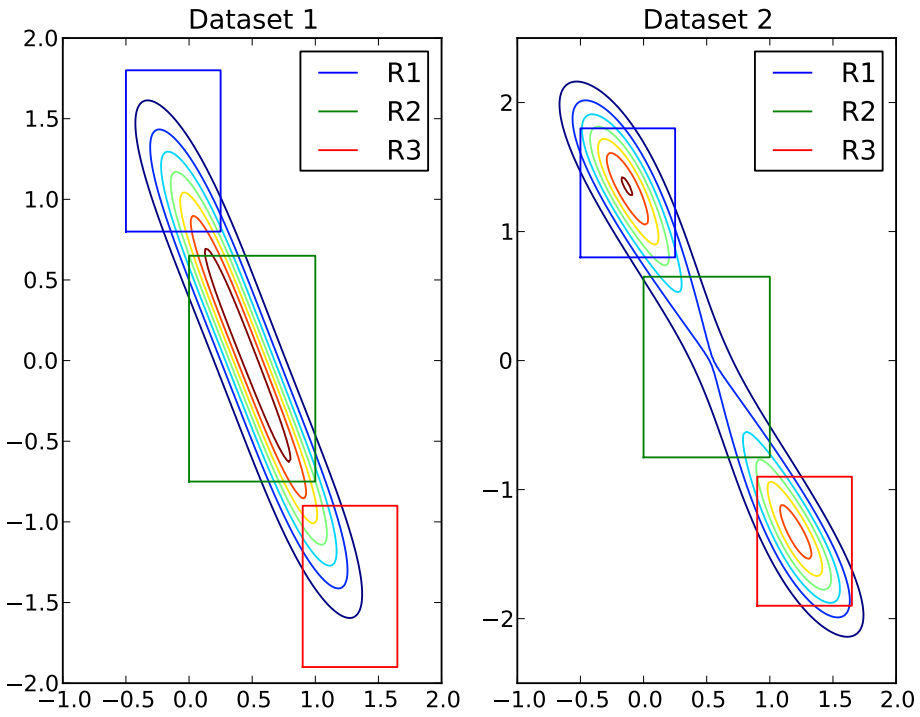
$$S.\widehat{E.(\Pr[R_i])} = \left(\frac{\widehat{\Pr[R_i]}(1 - \widehat{\Pr[R_i]})}{\left(\sum_{k=1}^{n} \mathbb{1}_{[-10,10]^2}(X_k)\right)/\tau} \right)^{1/2} \tag{16}$$

where τ is the autocorrelation time of the function $\mathbb{1}_{R_i}(X)$.

3.3 *Data*

For this experiment I use two artificially generated datasets (random seed = 17 and random seed = 19) according to the toy model both with $n = 100$. The first dataset, D_1, results in a unimodal posterior while the second, D_2 results in a bimodal posterior. The unimodal posterior is centered on the prior but the peak is elongated by the likelihood between the two modes we expect. By construction, both exhibit strong negative correlation between the θ's. The posteriors are plotted below with regions R_i indicated by the rectangles. The three test regions are chosen along the length of the distribution such that the mode(s) should be accounted for and, in the case of multiple modes, separated by the regions. You can see the posteriors for each distribution with the three regions identified in figure 1.

Figure 1: Posterior Distributions



3.4 *Algorithm Specifications*

All algorithms generate 10^6 draws. As a baseline algorithm I use an LMC proposal move with an MH step to ensure detailed balance is satisfied. I run the LMC sampler with several settings for ϵ (as defined in (2)). The SGLD sampler is run with several batch sizes, n, and for several annealing schedules for ϵ. I use the polynomial schedule defined in [7]: $a(b+t)^{-\gamma}$ but change the maximum and minimum ϵ. This polynomial schedule meets the two following criteria that are required for SGD convergence to a mode [2]:

$$\sum_{t=1}^{\infty} \epsilon_t > \infty\,, \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty\,.$$

Sampling from the data for each batch is done without replacement [6]. For SGLD, I test the following batch sizes: $n = 1, 2, 4, 8, 16$.

4 Results

The results are summarized in figures 2 to 7 that show the true statistic, estimated statistic and error bars for

$$\Pr_{D_j}[R_i]\,,\ i = 1, 2, 3 \text{ and } j = 1, 2$$

The first figure corresponds to the LMC sampler with the MH step while the next 6 correspond to the different batch sizes. The estimated statistic is shown for different annealing schedules for SGLD and different fixed step sizes for LMC. The figures for batch sizes > 1 are in the appendix.

Considering our exact LMC sampler first in figure 2 we see that for nearly every region and step size the true statistic falls within the error bars of our estimates. This is what we wanted to see. Given the narrow geometry and possibly low conductance of our posteriors (especiallin dataset D_2), these are the types of results one would expect and should be fairly satisfied with. The key observation here is that the estimates seem to be reasonably stable with respect to the choice of ϵ and the estimator performs equally well for almost all the regions. What's pleasant about these results is that it manages to sample well from both modes in D_2 even with bad geometry (see the first and third plot in the second column).

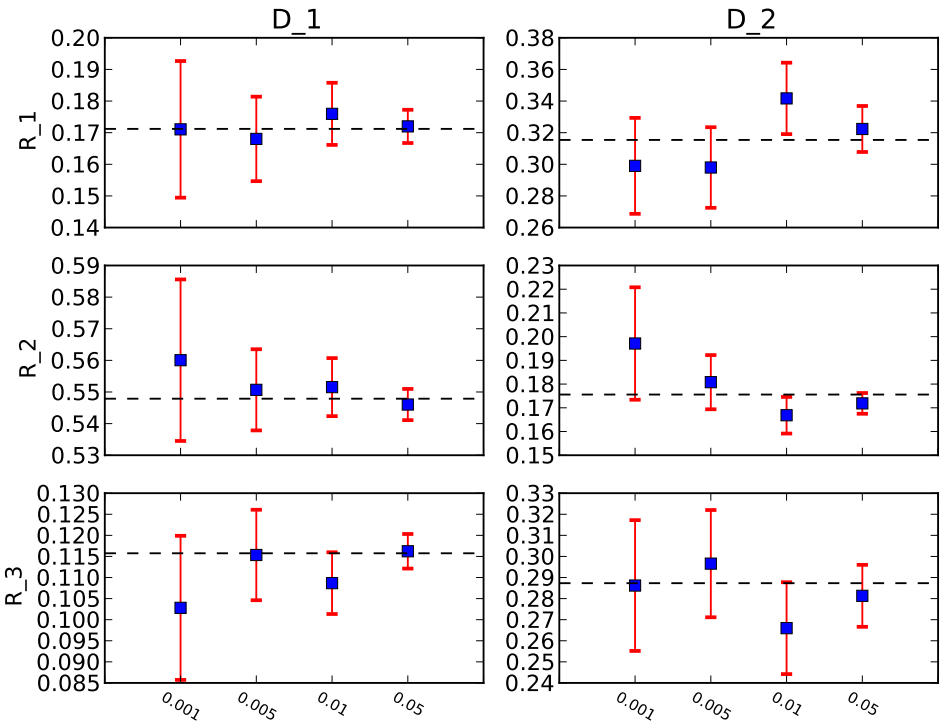
On the other hand, in figure 3 we see variable results from our SGLD with batch size of 1 (Note that the scales are not the same between the LMC plots and the SGLD plots). Few estimators catch the true statistic in their error bars and the estimates are very sensitive to annealing schedule. Some schedules are too slow and end up with zero estimates. This may be the result of the sample chain bouncing back and forth over the narrow axis of the distribution. As we might expect, the estimates seem to stabilize somewhat for $n = 16$ (at least for the unimodal distribution). But, this does not appear to be a monotonic effect with respect to n and for larger n we still see some bad annealing schedules. The performance for all SGLD methods seems to be a bit better for the unimodal distribution D_1.

[7] also suggests a method to re-weight the estimates based on ϵ_t to account for the annealing schedule. The intuition is that the effective sample size of a draw from the posterior is proportional to ϵ_t. The reweighted estimator $\hat{f}_w(X)$ can be written as

$$\hat{f}_w(X) = \frac{\sum_{i=1}^{N} \epsilon_t f(X_i)}{\sum_{i=1}^{N} \epsilon_t} .$$

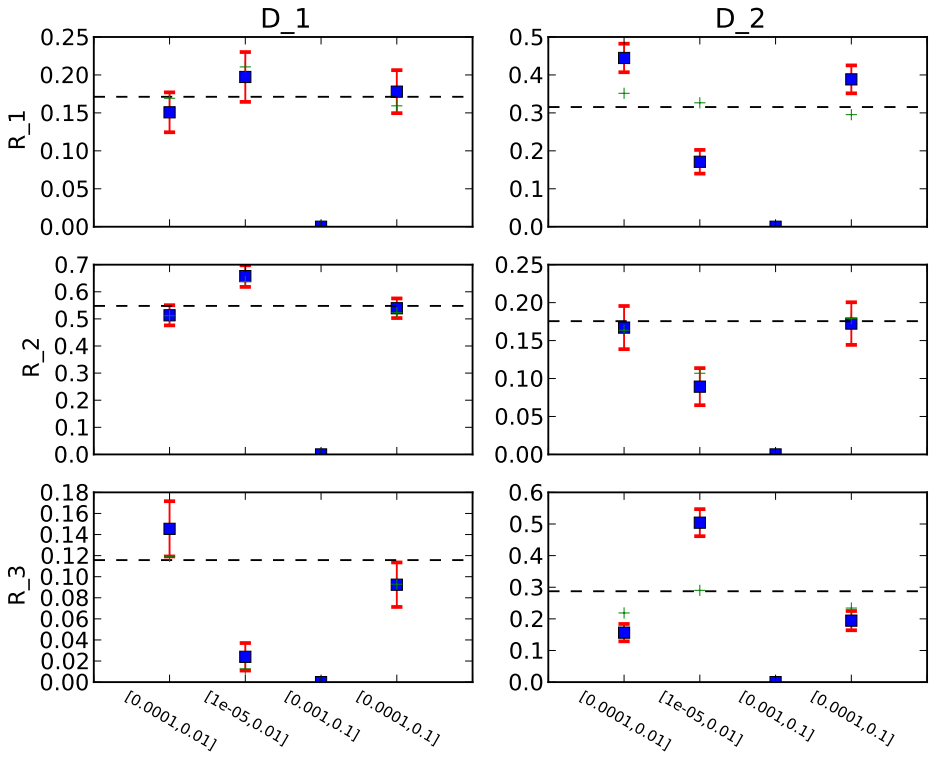
This has the unfortunate property of giving more weight to draws more likely to have been rejected. These estimates are represented by green crosses in the SGLD plots and one can see that its effect is not conclusive. Sometimes the estimator is significantly improved, sometimes there is no noticeable change, and sometimes it is worse. It appears that there may be a trade off between over emphasizing draws that were potentially rejected and eliminating some of the overrepresentation of states at the tail end of the chain when ϵ_t is very small.

Figure 2: LMC Error Analysis



Notes: Each row corresponds to the probability region being *Notes:* Each row corresponds to the probability region being estimated and each column corresponds to the dataset used to create the posterior distribution. (*Black dashed*) is the true statistic. The x-axis labels indicate the ϵ size for each run.

Figure 3: SGLD Error Analysis ($n = 1$)



Notes: Each row corresponds to the probability region being *Notes:* Each row corresponds to the probability region being estimated and each column corresponds to the dataset used to create the posterior distribution. When points are at zero it usually means the sampler did not converge properly and the region was never reached by the sampler. Each point represents a probability estimate for a region. The red bars indicate the estimated 95% error bands. (*Black dashed*) is the true statistic. (*Green crosses*) represent the reweighed estimators. Tuples on the x-axis are the minimum and maximum ϵ values of the annealing schedule.

5 Conclusion

The SGLD algorithm has some very obvious computational benefits but its performance is variable and very sensitive to annealing schedule and batch size. The hypothesis is that for large and redundant datasets using batches of an appropriate size, the stochastic gradient will be a good approximate of the true gradient. We see though in our error analysis that the discretization error for SGLD will lead to $O(N/n)$ expected number of rejections which seems to be important enough of an observation to be explored more thoroughly. This observation suggests that SGLD is inherently biased. Additionally, the fact that one needs to anneal the step size to zero causes the process to be highly autocorrelated. The experimental results suggest that the SGLD is in fact not an exact method but it does seem to perform incrementally better for more well behaved distributions.

[1] attempts to remedy some of these issues by invoking the so called Bayesian CLT that says when N is big the posterior is approximately normal with covariance equal to the inverse Fisher Information Matrix. Therefore, one should be able to construct an adaptive MCMC scheme that has this normal distribution as its invariant distribution. Such a scheme is adaptive in that it incrementally updates the estimate of the Fisher Information matrix.

What this preliminary paper suggests is that there is presently no free lunch when it comes to

sampling complicated and computationally intensive posteriors. An annealing step size may lead to good posterior maximization but increases autocorrelation time of the chain and causes it to be irreversible. Avoiding the rejection step which is crucial to guarantee the correctness of most MCMC methods leads to some unintended consequences and even when the probability of rejection goes to zero the number of rejections will be non-trivial and this will lead to biased estimates. This suggests that an interesting future direction of research might be in developing a theoretically sound approximate rejection step.

References

[1] Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *ICML*, 2012.

[2] L. Bottou. Large-scale machine learning with stochastic gradient descent, 2010.

[3] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2001.

[4] R. M. Neal. Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2010.

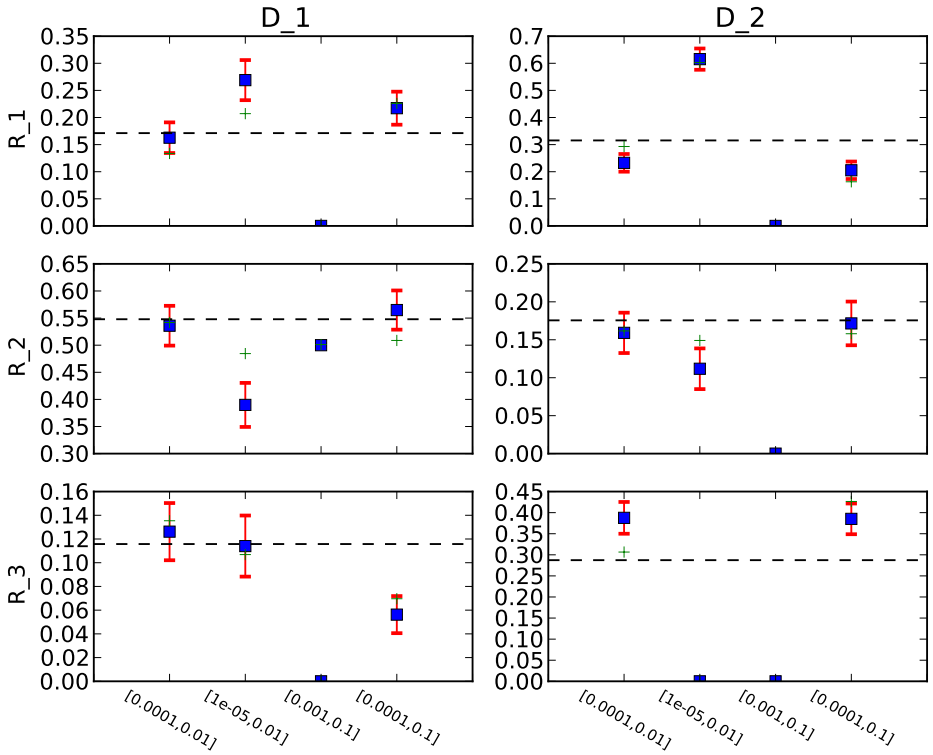
[5] Radford M. Neal. Probabilistic inference using markov chain monte carlo methods, 1993.

[6] F. Niu, B. Recht, C. Ré, and S. J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent, 2011.

[7] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

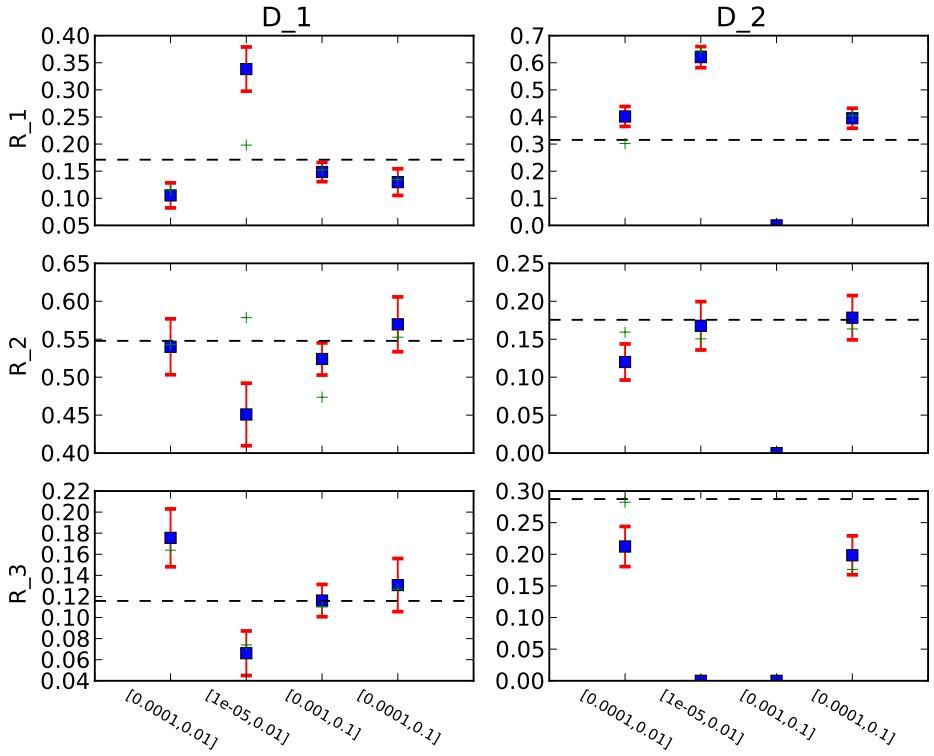
6 Appendix

Figure 4: SGLD Error Analysis ($n = 2$)



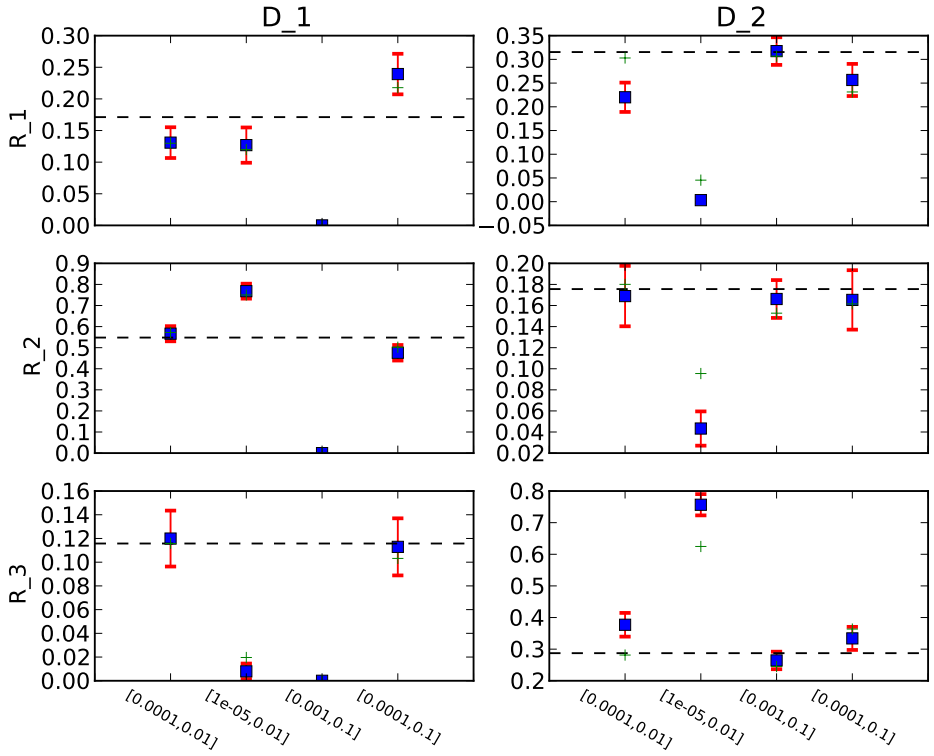
Notes: Each row corresponds to the probability region being *Notes:* Each row corresponds to the probability region being estimated and each column corresponds to the dataset used to create the posterior distribution. When points are at zero it usually means the sampler did not converge properly and the region was never reached by the sampler. Each point represents a probability estimate for a region. The red bars indicate the estimated 95% error bands. (*Black dashed*) is the true statistic. (*Green crosses*) represent the reweighed estimators. Tuples on the x-axis are the minimum and maximum ϵ values of the annealing schedule.

Figure 5: SGLD Error Analysis ($n = 4$)



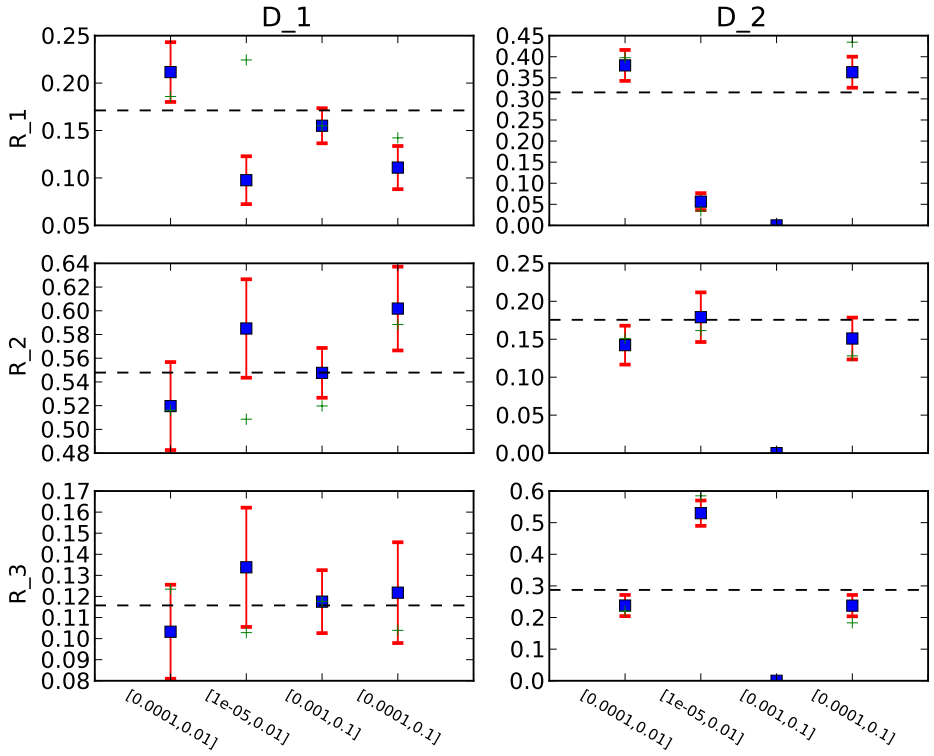
Notes: Each row corresponds to the probability region being *Notes:* Each row corresponds to the probability region being estimated and each column corresponds to the dataset used to create the posterior distribution. When points are at zero it usually means the sampler did not converge properly and the region was never reached by the sampler. Each point represents a probability estimate for a region. The red bars indicate the estimated 95% error bands. (*Black dashed*) is the true statistic. (*Green crosses*) represent the reweighed estimators. Tuples on the x-axis are the minimum and maximum ϵ values of the annealing schedule.

Figure 6: SGLD Error Analysis ($n = 8$)



Notes: Each row corresponds to the probability region being *Notes:* Each row corresponds to the probability region being estimated and each column corresponds to the dataset used to create the posterior distribution. When points are at zero it usually means the sampler did not converge properly and the region was never reached by the sampler. Each point represents a probability estimate for a region. The red bars indicate the estimated 95% error bands. (*Black dashed*) is the true statistic. (*Green crosses*) represent the reweighed estimators. Tuples on the x-axis are the minimum and maximum ϵ values of the annealing schedule.

Figure 7: SGLD Error Analysis ($n = 16$)



Notes: Each row corresponds to the probability region being *Notes:* Each row corresponds to the probability region being estimated and each column corresponds to the dataset used to create the posterior distribution. When points are at zero it usually means the sampler did not converge properly and the region was never reached by the sampler. Each point represents a probability estimate for a region. The red bars indicate the estimated 95% error bands. (*Black dashed*) is the true statistic. (*Green crosses*) represent the reweighed estimators. Tuples on the x-axis are the minimum and maximum ϵ values of the annealing schedule.