# Supplementary Material
# Xing, Nicholls and Lee, ICML 2019
# Calibrated Approximate Bayesian Inference

May 13, 2019

## 1 A toy example

We use a toy example similar to Lee et al (2018) to compare the performance of Algorithm 2 in our paper (we refer to this as the AIS sampler) and the IS sampler in Lee et al (2018) as the dimension of the observation $y_{obs}$ varies. It is inevitable that there exist regimes where AIS outperforms IS (on larger harder problems typically). The IS sampler breaks as the dimension of $y_{obs}$ grows and we take poor importance proposal distributions. AIS uses the same initialising distribution as IS but uses a sequential procedure to march the particle distribution onto the target.

The parameter $\phi$ is a draw from the prior $\phi \sim \mathcal{N}(0,1)$. The data are iid observations $\{y_1, ..., y_d\} \sim \mathcal{N}(\phi, 1)$. The exact posterior is

$$\pi(\phi|y_1, ..., y_d) \sim \mathcal{N}\left(\frac{\sum_{i=1}^d y_i}{d+1}, \frac{1}{d+1}\right).$$

Following Lee et al (2018), who consider this setup with $d = 1$, consider an approximation in which we replace the exact likelihood with a tempered likelihood $\tilde{p}(y|\phi) = \mathcal{N}(y|\phi)^v$ for some $v > 0$. The corresponding approximate posterior

$$\tilde{\pi}(\phi|y_1, ..., y_d) \propto \mathcal{N}(0,1)\prod_{i=1}^d \mathcal{N}(y_i|\phi)^v,$$

that is,

$$\tilde{\pi}(\phi|y_1, ..., y_d) = \mathcal{N}\left(\phi; \frac{v\sum_{i=1}^d y_i}{vd+1}, \frac{1}{vd+1}\right).$$

When $v = 1$ the approximate posterior is exact.

Let $\tilde{C}_Y$ be a level $\alpha$ credible interval computed for the approximate posterior. Since both the exact and approximate posterior are normally distributed, the

true value of the operational coverage $b(y) = \Pr(\phi \in \tilde{C}_Y | Y = y)$ can be exactly computed. In this example we set $\alpha = 0.95$

Consider estimating $b(y)$ using IS and AIS samplers. We start from a scalar observation $y$ (i.e. $d = 1$). Both algorithms work well when $0 < v < 2$, which agrees with results in Lee et al (2018) who prove (for the case where $d = 1$) that the weight variance is finite. However, the performance of the IS sampler declines for $v > 2$ since in this case the approximate likelihood $\tilde{p}(y|\phi)$, which is the importance sampling proposal distribution, is more concentrated than the true likelihood. In Figure 1, we set $v = 2$ and estimate $b(y)$ at 100 equidistant points $y$s over the interval $(-3, 3)$. We can see that $b_{IS}(y)$, the IS sampler estimate of the true coverage, has significantly higher variance, while $b^{AIS}(y)$ performs much better and is closer to the true values.

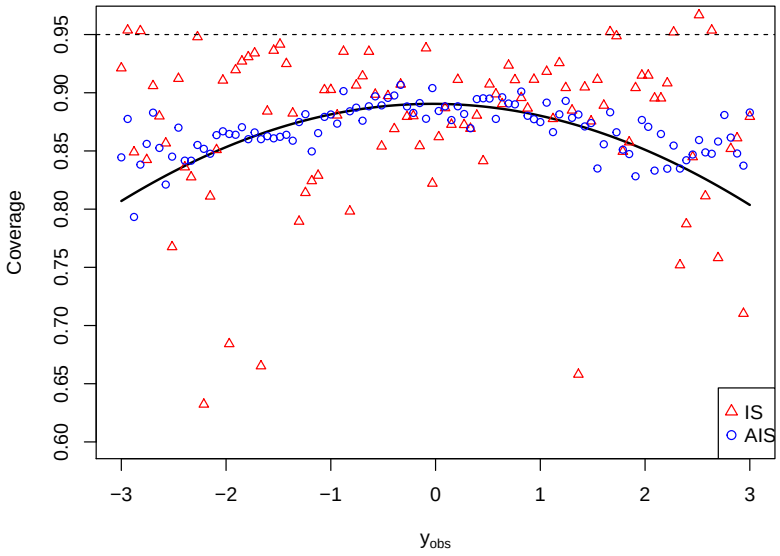


Figure 1: $y$ vs Coverage $b(y)$, the solid line corresponds to the true coverage $b(y)$, the IS and AIS estimate are represented by red and blue points respectively. The dashed line is the nominal coverage $\alpha = 0.95$

We extend the comparison to higher dimensions ($d > 1$). For each $d = 3, 6, 9, .., 30$, we sample $N_{reps} = 100$ synthetic parameters $\{\phi_i\}_{i=1}^{N_{reps}}$, from the prior distribution and $d$-component data vectors $y^{(i)} = \{y_1^{(i)}, ..., y_d^{(i)}\} \overset{iid}{\sim} \mathcal{N}(\phi_i, 1)$ for $i = 1, ..., N_{reps}$ as synthetic observations. We compute the corresponding estimate $b_{IS}(y^{(i)})$ and $b_{AIS}(y^{(i)})$ and the associated Effective Sample Size (ESS) values $ESS_{IS}^{(i)}$ and $ESS_{AIS}^{(i)}$ for particle weights using the two algorithms for each

$i = 1, ..., N_{reps}$ with $N_{AIS} = N_{IS} = 1000$ particles, then report the MSE $\bar{R}_{IS} = \frac{1}{N}\sum_{i=1}^{N_{reps}}(b_{IS}(y^{(i)}) - b(y^{(i)}))^2$ and $\bar{R}_{AIS} = \frac{1}{N_{reps}}\sum_{i=1}^{N_{reps}}(b_{AIS}(y^{(i)}) - b(y^{(i)}))^2$, and the corresponding average effective sample sizes.

We set $v = 5$, well above the safety zone given by Lee et al (2018). The approximate posterior is heavily under-dispersed with respect to the true posterior here and indeed, this choice (ie $v = 5$) causes IS to fail, in the sense that the ESS is often consistent with an ESS equal one. We report the naive average ESS for IS, which is about 35 at $N_{IS} = 1000$, but this cannot be trusted. We showed this by doubling the number of particles $N_{IS} \leftarrow 2N_{IS}$ to 2000. The ESS should (approximately) double if it is being reliably estimated. We find it increases only slightly. In contrast doubling $N_{AIS} \leftarrow 2N_{AIS}$ very nicely doubles $ESS_{AIS}$.

For a more extreme example, we repeated the simulation with $\nu = 10$ and $d = 100$ with $N_{AIS} = N_{IS} = 1000$ and 2000 particles. For 1000 particles, we find averages $ESS_{IS} = 35$ (again) and $ESS_{AIS} = 732$. When we double the number of particles to 2000, we get $ESS_{IS} = 44$ and $ESS_{AIS} = 1323$. Doubling the number of particles doubles $ESS_{AIS}$ but does not improve $ESS_{IS}$ a great deal, reflecting the fact that the IS sampler fails at these extreme $v$ and $d$ values, and the true $ESS_{IS} \sim 1$ in this case.

We do not consider computation times here - we look at efficiency in the next section using real data. Our point here is that IS has completely failed so computation times are irrelevant.

Results are reported in Table 1. We can see that as $d$ increases, the AIS sampler outperforms the IS sampler in both MSE and average ESS. This supports the effectiveness of our algorithm in high dimensional cases. Note that the MSE for IS, though larger than AIS, is still reasonable, as the breakdown impacts variance, and the tail behavior of estimates, though the mean is reasonably stable.

| $d$ | $\bar{R}_{IS}$ | $\bar{R}_{AIS}$ | $ESS_{IS}$ | $ESS_{IS}$ | $ESS_{AIS}$ | $ESS_{AIS}$ |
|---|---|---|---|---|---|---|
| 6 | 4.10e-02 | **3.11e-03** | 36 | 55 | **238** | **525** |
| 12 | 3.72e-02 | **1.69e-03** | 33 | 53 | **498** | **936** |
| 18 | 3.77e-02 | **3.39e-03** | 35 | 46 | **649** | **1258** |
| 24 | 4.12e-02 | **5.05e-03** | 34 | 50 | **726** | **1442** |
| | $\mathrm{N}_{IS} = 1000$ | $\mathbf{N}_{AIS} = 1000$ | $\mathrm{N}_{IS} = 1000$ | $\mathrm{N}_{IS} = 2000$ | $\mathbf{N}_{AIS} = 1000$ | $\mathbf{N}_{AIS} = 2000$ |

Table 1: The average MSE and ESS of both Algorithms over $N_{reps} = 100$ repetitions under different dimensions $d$ using $N_{AIS} = N_{IS} = 1000, 2000$. We report the naive ESS for IS. This is around 35 for $N_{IS} = 1000$ but is an upward biased estimate (see text) and the actual value is consistent with one. This is evidenced by re-estimation of the ESS at $N_{IS} = 2000$

## 2 Remark on Model Misspecification

Model misspecification does not enter the definition of coverage. The coverage probability for a procedure is the probability a credible set computed from data $y$ covers a parameter $\phi$ when $\phi$ and $y$ are a realisation of the generative model $\pi(\phi)p(y|\phi)$. This is how the level of a credible set is defined in Bayesian inference. It would certainly be interesting to know if we are covering nature's true parameter. However we estimate the change in coverage as we move from the exact to the approximate posterior, not the change coverage as we move from nature's generative model to the misspecified model.

However, model misspecification does have an impact on the difficulty of forming reliable coverage estimates: if the model is misspecified, then the data is an outlier, because the data is located in a part of data-space we don't often visit with our simulated $\{\phi_i, y_i\}$ pairs from the generative model. This makes estimation harder. In the regression approach, large extrapolation may lead to unstable estimates.

## 3 Computational Efficiency, a comparison

Both the IS and AIS sampler sample from the observation model $p(\cdot|\theta)$ and approximate posterior $\hat{\pi}(\cdot|y)$. If $T_1$ is the time to sample synthetic data $y' \sim p(y'|\phi)$ (often fast), and $T_2$ is the time to sample $\theta \sim \hat{\pi}(\theta|y')$, the approximate posterior (often slow), then the IS sampler costs about $M(T_1 + T_2)$ for M particles and AIS costs about $MJT_1 + T_2$ ($J$ intermediate AIS steps, and we only need one set of approximate posterior MCMC run for initialisation). For fixed $M$, simulation studies show that AIS has a much bigger ESS. However there exist problems where $T_1$ is slow and $T_2$ is fast. For example, in Section 4 of the main paper, the approximate posterior for the Ising model can be evaluated exactly (so $T_2$ is effectively 0), but sampling from the Ising model is time-consuming ($T_1$ is large). This means there exist problems where IS beats AIS, but not typically big hard problems (where the ESS of AIS sampler is much greater than the ESS of IS sampler and $T_2$ is big, like the random effect partition model in Section 5). Even on the Ising model where everything is in the IS sampler's favor, we show below that the IS sampler fails on large lattices where ESS estimates become unreliable (like harmonic mean) while AIS gives reliable ESS estimates even on large lattices.

To compare the effectiveness of the IS and AIS sampler, we rerun the Ising model simulation using both algorithms on a $N_I \times N_I$ lattice where the size of lattice $N_I = 25, 50, 75, 100, 125$. We initialise both algorithms with $M = 500$ particles. We report the effective sample size (per CPU second) for both algorithms for different sizes $d$ in Table 2. In Figures 2 and 3 we see that although the IS sampler is more efficient than the AIS sampler on smaller lattices, the performance of AIS sampler is much better when the size of lattice is greater than about $N_I = 75$. Actually the comparison is already kind to IS. Once the ESS becomes small it becomes hard to measure and ESS estimates themselves

become very noisy. The IS ESS estimates at large $N_I$ are probably consistent with one, so although the IS efficiency curve appears to flatten in Figure 2, the truth is very likely to be that it plummets out of the frame, as it is very unlikely to remain constant at increasing $N_I$. Larger ESS-values (like those seen for AIS) are much more reliably estimated.

| $d$ | $ESS_{IS}$ | $ESS_{AIS}$ | $ESS_{IS}$/s | $ESS_{AIS}/s$ |
|---|---|---|---|---|
| 25 | 7.468e+01 | **2.039e+02** | 7.732e-03 | **7.272e-03** |
| 50 | 9.095e+01 | **2.085e+02** | 9.332e-03 | **7.041e-03** |
| 75 | 2.129e+01 | **1.677e+02** | 1.676e-03 | **5.367e-03** |
| 100 | 2.675e+01 | **1.203e+02** | 2.306e-03 | **3.642e-03** |
| 125 | 1.967e+01 | **1.120e+02** | 1.506e-03 | **3.230e-03** |

Table 2: Effective sample sizes for for different lattice size. The first two columns are the actual ESS of the IS and AIS sampler, the third and forth columns are the ESS per CPU second figures for the two algorithms. The ESS values for IS above about $N_I = 75$ are consistent with one, so the efficiencies reported above for IS are overestimates.

# 4 Credible set for example in Section 5

Our purpose in the main paper is to demonstrate that we can reliably calibrate the coverage probability of an HPD credible set over partitions. The credible set itself would be of interest in the application, but is only of secondary interest to us.

In Table 3 we give the credible set for the example considered in the main paper Section 5. This is a partition of 12 levels of a treatment variable in a complete design with four block variables and five covariates in all. The favored partition $(1, 2, 8, 12, 10, 5), (11, 4, 7, 9, 3, 6)$ splits the levels into two groups.

# 5 BART on the Ising model example

We run BART on the Ising model example in the main paper Section 4. We fit two BART models using the natural sufficient statistics $S(y) = f(y, E_f)$ and the raw $200 \times 200$ binary image as input. The fitted curve is reported in Figure 4. We see that the fitted curve of BART trained by the natural sufficient statistics $S(y)$ (left) agrees with the fitted curve of BART trained by the raw image (which can be seen as a $40000 \times 1$ binary vector) . This learned the similar pattern and reproduces a curve similar to the left with greater uncertainty (right).

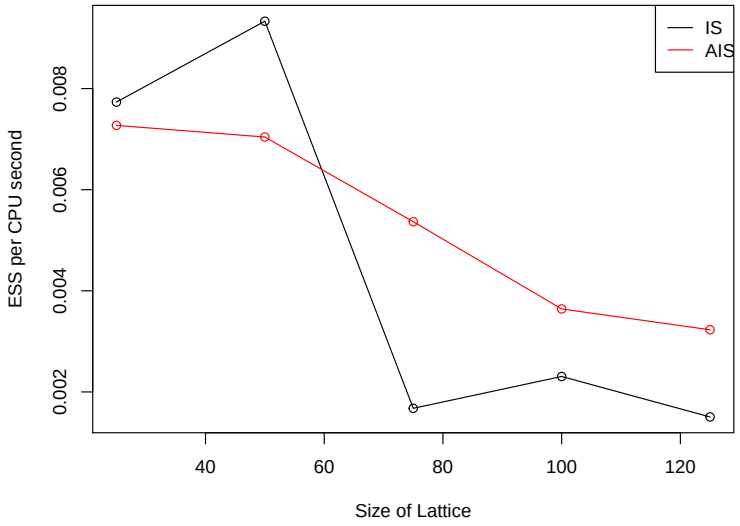


Figure 2: ESS per CPU second (efficiency) for both algorithms. IS efficiency estimates for Lattice sizes above 75 are likely to be over-estimates.

Table 3: A 95% approximate credible set for partition $S$ using $\tilde{\pi}(S|y_{obs})$. Partitions on the first column are sorted by their posterior probability in a decreasing order. The second column records the cumulative sum of posterior probabilities (i.e. $\tilde{G}(S|y_{obs})$ is the CDF of $\tilde{\pi}(S|y_{obs})$).

| Partition $S$ | $\tilde{G}(S\|y_{obs})$ |
|---|---|
| (1,2,8,12,10,5),(11,4,7,9,3,6) | 0.13 |
| (1,8,12),(2,10,5),(11,4,9,6),(7,3) | 0.24 |
| (1,2,8,12,10,5),(11,4,6),(7,9,3) | 0.28 |
| (1,2,8,12,10,5),(11,4,9,6),(7,3) | 0.31 |
| ⋮ | ⋮ |
| (1,8,11,4,9,6),(2,7,3),(12,10),(5) | 0.95 |

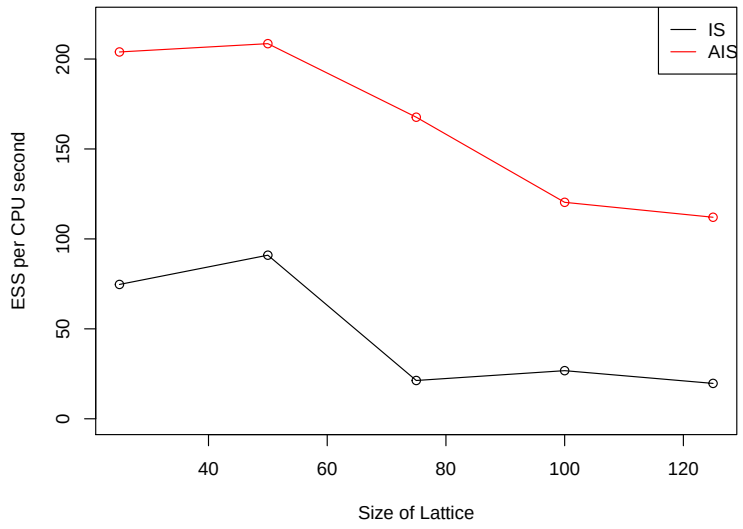


Figure 3: Actual ESS for both algorithms. The ESS values for IS above about $N_I = 75$ are consistent with one.

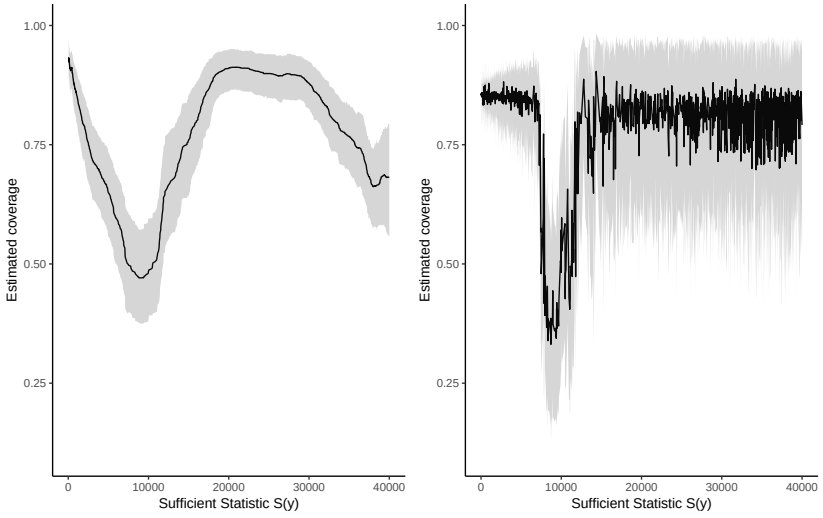


Figure 4: Sufficient statistics $S(y)$ vs Estimated coverage. Left: BART trained by the sufficient statistics. Right: BART trained by the full $200 \times 200$ image. Grey band indicates the 95% credible interval.