Sampling using adaptive regenerative processes: supplementary material

HECTOR MCKIMM$^{1,a}$, ANDI WANG$^{1,b}$, MURRAY POLLOCK$^{2,e}$, CHRISTIAN ROBERT$^{1,c}$ and GARETH ROBERTS$^{1,d}$

$^1$Department of Statistics, University of Warwick, Coventry, United Kingdom, $^a$Hector.McKimm@warwick.ac.uk, $^b$andi.wang@warwick.ac.uk, $^c$C.A.M.Robert@warwick.ac.uk, $^d$Gareth.O.Roberts@warwick.ac.uk
$^2$School of Mathematics, Newcastle University, Newcastle-upon-Tyne, United Kingdom, $^e$murray.pollock@newcastle.ac.uk

Keywords: Adaptive algorithm; Markov process; MCMC; normalizing constant; regeneration distribution; Restore sampler; sampling; simulation

Appendix A: Output

When output times are fixed, let \( \{t_1, t_2, \ldots \} \) be an evenly spaced mesh of times, with \( t_i = i\Delta \) for \( i = 1, 2, \ldots \) and \( \Delta > 0 \) some constant. When output times are random, let \( \{t_1, t_2, \ldots \} \) be the events of a homogeneous Poisson process with rate \( \Lambda_0 > 0 \). In either case, the output of the process is \( \{X_{t_1}, X_{t_2}, \ldots \} \). Suppose there are \( n \) output states, then we estimate expectations using the unbiased approximation:

\[
\pi[f] \approx \frac{1}{n} \sum_{i=1}^n f(X_{t_i}).
\]

Algorithmically, there is little difference between using fixed and random output times. The memoryless property of Poisson processes allows one to generate the next potential regeneration and output events, \( \tilde{\tau} \) and \( s \), simulate the process forward in time by \( \tilde{\tau} \lor s \), then discard both \( \tilde{\tau} \) and \( s \). When using a fixed mesh of times, the memoryless property no longer applies, so one must keep track of the times of the next output and potential regeneration events.

Appendix B: Pre-transformation of the target distribution

In the multi-dimensional setting, the Brownian Motion Restore sampler is far more efficient at sampling target distributions for which the correlation between variables is small. Rate \( \tilde{\kappa} \) is more symmetrical for target distributions \( \pi \) with near-symmetrical covariance matrices. Since the Markov transition kernel for Brownian motion over a finite period of time is symmetrical, local moves are better suited to near-symmetrical target distributions.

More generally, the parameterization of \( \pi \) has a large effect on Bayesian methods (Hills and Smith, 1992). In practice, we recommend making a transformation so that the transformed target distribution has mean close to zero and covariance matrix close to the identity. Suppose we have \( X \sim N(m, \Sigma) \) and that \( \Sigma = V\Lambda V^T \) for \( V \) a matrix with columns the eigenvectors of \( \Sigma \) and the corresponding eigenvalues forming a diagonal matrix \( \Lambda \). Then for \( X' \sim N(0, I_d) \), we have \( X = \Sigma^{1/2} X' + m \), where \( \Sigma^{1/2} = V\Lambda^{1/2} \) and \( \Lambda^{1/2} \) is a diagonal matrix with entries the square roots of the eigenvalues of \( \Sigma \). It follows that when \( X \) is roughly Gaussian, with mean and covariance matrix \( m \) and \( \Sigma \), letting \( \Sigma^{-1/2} = (\Sigma^{1/2})^{-1} \).
transformed variable $X' = \Sigma^{-1/2}(X - m)$ should be close to an isotropic Gaussian. By the change of variables formula:
\[ \pi_{X'}(x') = \pi_X(x) \frac{dx}{dx'} = \pi_X(\Sigma^{1/2}x' + m)|\Sigma^{1/2}|. \]

In computing the gradient and Laplacian of the energy of the transformed distribution, one must use the chain-rule to take into account the matrix $\Sigma^{1/2}$. Samples obtained from $\pi_{X'}$ may be transformed to have distribution $\pi_X$.

In most of the examples presented in this paper, the target distribution undergoes a pre-transformation as above, with $m$ and $\Sigma$ estimated by a Laplace approximation. For notational simplicity, we will continue to refer to sampling random variable $\pi X$ with distribution $\pi$, even when in actual fact we are sampling the transformed distribution $\pi_{X'}$ corresponding to transformed variable $X'$. We make the Laplace approximation using the “optim” function in R (R Core Team, 2021), which uses numerical methods to find the mode of $\pi$ and the Hessian matrix of log $\pi$ at the mode.

Appendix C: Logistic Regression Model of Breast Cancer

The data (Mangasarian and Wolberg, 1990) was obtained from the University of Wisconsin Hospitals, Madison. The response is whether the breast mass is benign or malignant. Predictors are features of an image of the breast mass. The model has dimension $d = 10$. We used a Gaussian product prior with variance $\sigma^2 = 400$. Following Gelman et al. (2008), we scaled the data so that response variables were defined on $(-1, 1)$, non-binary predictors had mean 0 and standard deviation 0.5, while binary predictors had mean 0 and range 1. The posterior distribution was transformed based on its Laplace approximation, as described by Appendix B.

As described in Section 5.1, we simulated 100 samples paths of an Adaptive Restore process, with stationary distribution the posterior distribution of the logistic regression model of breast cancer. Each sample path was generated for a simulation time of $T = 10^6$, then the samples divided into 10 batches. Figure 1 shows boxplots of batch estimates of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2]$ for $i = 1, \ldots, 10$.

Figure 2 displays the MSE of estimates of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2]$, for $i = 1, \ldots, 10$, for four types of Restore processes (as described in section 5.1). The burn-in time was $b = 3 \times 10^5$ and the total simulation time $T = 4 \times 10^5$.

Appendix D: Justification for the minimal regeneration distribution

Throughout the paper, we consider targeting the minimal regeneration distribution within our Adaptive Restore algorithm. The reason for this is that we wish to minimise the computational cost per unit time of the algorithm. In subsection 2.4.3, we stated that an advantage of using $k^+$ for Restore simulation is that to ensure $\mathbb{P}[k^+ < \mathcal{K}^+] < 1 - \epsilon$ is satisfied, for $\epsilon > 0$ a small constant (e.g. $\epsilon = 0.001$). $\mathcal{K}^+$ scales logarithmically with dimension $d$. To see this, consider $X \sim \mathcal{N}(0, I)$. Then,
\[ \mathbb{P}[k^+(X) < \mathcal{K}^+] = \mathbb{P}[0.5(x^T x - d) < \mathcal{K}^+] = \mathbb{P}\left[ \sum_{i=1}^{d} x_i^2 < 2\mathcal{K}^+ + d \right] = \mathbb{P}[Q < 2\mathcal{K}^+ + d], \]
for $Q \sim \chi^2_d$. Figure 3 shows $\mathcal{K}^+$ so that $\mathbb{P}[k^+(X) < \mathcal{K}^+] < 1 - \epsilon$ for $\epsilon = 0.01, 0.001, 0.0001$. Thus for a 100-dimensional Gaussian target distribution, the truncation level $\mathcal{K}^+ = 30$ would likely be appropriate.

As a caveat, some functions, such as $f(x) = x^T x$, are very sensitive to $\mathcal{K}^+$ so an even more conservative
Figure 1: Boxplots of batch estimates of $\mathbb{E}[X_i]; i = 1, \ldots, 10$ of the posterior of a Logistic Regression model of breast cancer, computed using 100 samples paths of Adaptive Restore processes.
D JUSTIFICATION FOR THE MINIMAL REGENERATION DISTRIBUTION

Figure 2: MSE of estimates of $E[X_i]; i = 1, \ldots, 10$, for different types of Restore processes. Circles and crosses ($\times$) show MSEs for Adaptive Restore processes, with $\mu_0 \equiv N(0, I)$ and $\mu_0 \equiv \mu_{N_0}$ respectively. Triangles and crosses ($+$) show MSEs for Standard Restore processes, with $\mu \equiv N(0, I)$ and $\mu \equiv N(0, 1.5I)$ respectively.
Figure 3: For $\pi \equiv N(0, I)$, plot of $K^+$ so that $P[k^+(X) < K^+] = 1 - \epsilon; \epsilon = 0.01, 0.001, 0.0001$.

choice of $K^+$ might be necessary. Furthermore, $k$ may prove impossible to derive in realistic situations.

However, reducing regenerations could conceivably make Markov chains mixing slower. In this section, we consider a stylised example in a formal asymptotic analysis for large dimensions where we demonstrate that the minimal regeneration distribution in fact leads to $O(1)$ convergence time per unit regeneration.

Suppose that we are targeting $\pi = N_d(0, I_d)$ and for a random variable $X$ on $\mathbb{R}^d$ set $R = |X|$. We consider minimal regeneration Restore with Brownian motion local dynamics. Notice that by simple calculations we have that

$$ L^+\pi(x) = |x|^2 - d $$

and

$$ \tilde{k}(x) = \left( L^+\pi \right)_{+} = \frac{(|x|^2 - d)_+}{2}.$$

Furthermore we have that

$$ C^+\mu^*(x) = (L^+\pi)_- = \frac{(d - |x|^2)_+ e^{-|x|^2/2}}{(2\pi)^{d/2}}.$$

Now given the spherical symmetry of this example, we marginalise to the radial component:

$$ C^+\tilde{\mu}_R(r) = \frac{r^{d-1}(d - r^2)_+ e^{-r^2/2}}{(2\pi)^{d/2}}.$$

Now to understand how this minimal regeneration Restore behaves as $d \to \infty$, we must rescale $R$. Note that for large $d$, $R \approx N(d^{1/2}, 1/2)$. Therefore set $Y := R - d^{1/2}$ and translate the regeneration distribution (and rate) to the $Y$ space.

$$ C^+\tilde{\mu}_Y(y) = \frac{(y + d^{1/2})^{d-1}(d - (y + d^{1/2})^2)_+ e^{-(y+d^{1/2})^2/2}}{(2\pi)^{d/2}}, \text{ for } y > 0, $$

$$ \propto \left( 1 + \frac{y}{d^{1/2}} \right)^{d-1} \left( -2y - y^2/d^{1/2} \right)_+ e^{-y^2/2} e^{-yd^{1/2}}.$$
Now the local dynamics of $R$ are well-known to follow the Bessel$(d)$ process, so that for large $d$ we have that the local dynamics for $Y$ are
\[
\begin{align*}
\frac{dY_t}{dt} &= dB_t + \frac{d - 1}{2d^{1/2}} dt, \\
&\approx dB_t + \frac{d^{1/2}}{2} dt.
\end{align*}
\]

Moreover the regeneration rate is
\[
\bar{\kappa} = \frac{1}{2} (\|x\|^2 - d) = \frac{1}{2} (y^2 + 2yd^{1/2}) = \approx yd^{1/2}.
\]
Since the regeneration rate and local dynamics are both $O(d^{1/2})$, so further slowing down the process by a factor of $d^{1/2}$, we finally set $Z_t = Y_{d^{-1/2}t}$, and ignoring terms negligible for large $d$, the local dynamics reduces to
\[
\frac{dZ_t}{dt} \approx \frac{1}{2} dt,
\]
namely a constant velocity deterministic flow; the regeneration rate reduces to $z_+$, while the minimal regeneration density is
\[
\tilde{\mu}(z) \propto (z)_+ e^{-z^2}.
\]

For a formal proof of this final statement, see (Wang, 2020, Theorem 5.6.1).

**Remark 1.** Note that this non-reversible Markov process can be readily checked to permit $N(0, 1/2)$ as stationary distribution as expected, by showing that its generator applied to an arbitrary $L^2$ function has mean zero with respect to $N(0, 1/2)$.

Convergence of $Z$ is $O(1)$ in $d$-dimensions, as is its regeneration rate. Therefore the minimal regeneration distribution does indeed achieve $O(1)$ convergence per regeneration in high-dimensional contexts.

In contrast, consider a strategy of attempting to construct a regeneration distribution which is close to $\pi$. If this can be achieved exactly, the resulting Restore algorithm will work extremely well. Of course, however, this will not be achievable in practice. Suppose in the above example we instead manage to achieve a regeneration distribution
\[
\mu(x) \sim N(d, (1 + \epsilon)I_d).
\]

Now without repeating the entire analysis, $R$ concentrates like $(1 + \epsilon)^{1/2}d^{1/2}$, while the local dynamics propels the process it to even higher values and exponentially large regeneration rates. Moreover, the probability of regenerating a value of $Z < 1$ (where half its stationary mass lies) is exponentially small in $d$ (following from elementary estimates on the $\chi^2_d$ distribution). So regardless of how small $\epsilon$ is, convergence per regeneration is exponentially slow as a function of dimension.

Instead one might hope that using an *under*-dispersed regeneration distribution might be more robust. To investigate this we instead set
\[
\mu(x) \sim N(d, (1 - \epsilon)I_d).
\]
for some $\epsilon > 0$ and again propose to use Brownian motion local dynamics. In this case we see that

$$\kappa(x) = \|x\|^2 - d + Ce^{-\epsilon \|x\|^2/2},$$

for $\xi = \epsilon/(1 - \epsilon)$. This $\kappa$ is minimised at

$$\|x\|^2 = \frac{2\log(C\xi/2)}{\xi},$$

with minimised value $2/\xi - d + 2\log(C\xi/2)/\xi$. In order to make this non-negative we therefore require

$$C \geq \frac{2\exp(\xi d/2 - 1)}{\xi}.$$

Thus unless $\xi$ (and hence $\epsilon$) is smaller than $2/d$, $C$ will be forced to be exponentially large in $d$ and thus leading to exponential complexity of the algorithm once more.

As a result of these calculations, we see that the strategy of directly targeting $\pi$ as the regeneration distribution lacks robustness in its convergence.

The overall conclusion of these calculations is that targeting the minimal regeneration distribution has the best chance of breaking the worst effects of the curse of dimensionality, thus justifying the strategy we take in this paper.

**Appendix E: Further Experiments**

We include some further experiments, which serve to check the correctness of the implementation: a transformed Beta distribution and a multivariate Gaussian distribution.

**E.1. Transformed Beta Distribution**

We experiment with sampling from a distribution with density

$$\pi(x) = 6e^{2x}(e^x + 1)^{-4}.$$

This distribution is derived from the transformation of a Beta distribution. Consider $X' \sim \text{Beta}(2, 2)$, so that $\pi_{X'}(x') \propto x'(1-x')$ for $x' \in [0, 1]$. Let $X$ be defined by the logit transformation of $X'$, that is $X = \log\left(\frac{X'}{1-X'}\right)$, so that $X$ has support on the real line. The inverse of this transformation is $X' = e^X/(e^X + 1)$, so the Jacobian is $\frac{dx'}{dx} = \frac{e^x}{(e^x + 1)^2}$. Thus $X$ has density:

$$\pi(x) = 6\frac{e^x}{e^x + 1} \left(1 - \frac{e^x}{e^x + 1}\right) \frac{e^x}{(e^x + 1)^2} = \frac{6e^{2x}}{(e^x + 1)^4}.$$

We make no further transformation of the target. The partial regeneration rate is

$$\tilde{\kappa}(x) = \frac{4e^{2x} - 12e^x + 4}{2(e^x + 1)^2}.$$

It happens that $\tilde{\kappa}(x) < 2, \forall x \in \mathbb{R}$, which makes this distribution a useful test case, since an Adaptive Restore process may be efficiently simulated without any truncation of the regeneration rate. In addition, the first and second moments, 0 and $(\pi^2 - 6)/3$, may be computed analytically. Here, $K^- = 0.5$. 

Taking $\mu_0 \equiv \mathcal{N}(0, 1)$ we simulated 100 Adaptive Restore processes with Short-Term Memory, with parameters $T = 10^5$, $A_0 = 2$, $a = 10$, $n_{\text{cloud}} = 10^4$, $n_{\text{forget}} = 2$. For each sample path, the first half of the output states were burnt. 54 of the 100 estimates of the second moment were greater than the exact second moment; this indicates the processes have (approximately) converged to the correct invariant distribution.

E.2. Multivariate Gaussian distribution

We use an Adaptive Restore process with Short-Term Memory to sample a multivariate Gaussian distribution with dimension $d = 10$, mean $0.5 \times I_{10}$, variances $(0.92, 0.94, \ldots, 1.10)$ and pairwise covariances 0.5. The parameters used were $K^- = -5.05$ (3.s.f), $K^+ = 11.2$ (3.s.f), $A_0 = 1$, $a = 10$, $n_{\text{cloud}} = 10^4$, $n_{\text{forget}} = 2$ and $T = 2 \times 10^5$. The first half of the samples was removed and the remaining samples used to estimate $\mathbb{E}[X^T X]$. The MSE of the estimates was $5.32 \times 10^{-4}$ (3.s.f). Of the 100 estimates, 49 exceeded the true value of $\mathbb{E}[X^T X]$.

Appendix F: Hierarchical Model of Pump Failure

Figure 4 displays boxplots of batch estimates of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2]$; $i = 1, \ldots, 11$; for $\pi$ the posterior of a hierarchical model of pump failure. 100 samples paths of Adaptive Restore processes were simulated (as described in Section 5.2) and the samples from each path divided into 10 batches.

For the same posterior, Figure 5 displays MSEs of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2]$; $i = 1, \ldots, 11$; for different types of Restore processes, as described in section 5.2.

References

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Statist. 2 1360–1383. https://doi.org/10.1214/08-AOAS191

HILLS, S. E. and SMITH, A. F. M. (1992). Parameterization Issues in Bayesian Inference. In Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 227–246. Oxford University Press, Oxford, U.K.

MANGASARIAN, O. L. and WOLBERG, W. H. (1990). Cancer Diagnosis via Linear Programming. SIAM News 23.

R CORE TEAM. (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

WANG, A. Q. (2020). Theory of killing and regeneration in continuous-time Monte Carlo sampling. PhD thesis, University of Oxford.
(a) Estimates of $\mathbb{E}[X_i]$ for $i = 1, \ldots, 11$.

(b) Estimates of $\mathbb{E}[X_i^2]$ for $i = 1, \ldots, 11$.

Figure 4: Boxplots of batch estimates of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2]; i = 1, \ldots, 11$; for the posterior of the hierarchical model of pump failure; generated using 100 Adaptive Restore processes. A horizontal dashed line shows a very accurate approximation of the true value, computed using a long Markov chain.
Figure 5: MSEs of estimates of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i^2]$ for $i = 1, \ldots, 11$ for the posterior distribution of the hierarchical model of pump failure. The MSEs were computed by simulating 100 samples paths. Circles and crosses correspond to Adaptive Restore processes with $\mu_0 \equiv \mathcal{N}(0, I)$ and $\mu_0 \equiv \mu_N^*$ respectively. Triangles correspond to Standard Restore processes with $\mu \equiv \mathcal{N}(0, 3I)$. 