SMC2019 Hand In

Alma Andersson
almaan@kth.se

September 30, 2019

1 H.1. Importance Sampling Theory

a) **Given :** Proposal $q(x)$ and target $\pi(x) = \hat{\pi}(x)/Z$.
Objective : Show that $\hat{Z} = \frac{1}{N}\sum_i^N \frac{\hat{\pi}(x^i)}{q(x^i)}$ is an unbiased estimator of the normalizing constant
Solution : With $x_i \sim q(x)$ we have that

$$\mathbb{E}_q[\hat{Z}] = \mathbb{E}_q[\frac{1}{N}\sum_i^N \frac{\hat{\pi}(x^i)}{q(x^i)}] = \frac{1}{N}\sum_i^N \mathbb{E}_q[\frac{\hat{\pi}(x^i)}{q(x^i)}] = \mathbb{E}_q[\frac{\hat{\pi}(x)}{q(x)}] \tag{1}$$

Additionally

$$\mathbb{E}_q[\frac{\hat{\pi}(x)}{q(x)}] = \int \frac{\hat{\pi}(x)}{q(x)}q(x)dx = \int \hat{\pi}(x)dx = Z \tag{2}$$

Hence

$$\lim_{N\to\infty} |\frac{1}{N}\sum_i^N \frac{\hat{\pi}(x^i)}{q(x^i)} - Z| = 0 \tag{3}$$

Q.E.D.

b) **Given :** The IS (Importance Sampling) estimator of the normalization constant follows the CLT in eq 4

$$\sqrt{N}(\frac{\hat{Z}}{Z} - 1) \to \mathcal{N}(0, \int \frac{\pi(x)^2}{q(x)}dx - 1) \tag{4}$$

Objective : (i) Find the value of $\gamma > 0$ that minimizes the asymptotic variance when the proposal $q(x)$ as given in eq 5 is used for a standard Gaussian target distribution ($\pi(x)$). (ii) Show how the approximation of $\hat{Z}$ evolves over time.

$$q(x) = \frac{\gamma}{\pi \cdot (\gamma^2 + x^2)}, x \in \mathbb{R} \tag{5}$$

Start by defining a function $v(\gamma)$ that describes the asymptotic variance as a function of γ

$$v(\gamma) = \int_0^\infty \frac{\pi(x)^2}{q(x)}dx - 1 = \int_0^\infty \frac{1}{2\gamma}\exp(-x^2)(\gamma^2 + x^2)dx - 1 \tag{6}$$

The derivative of v w.r.t. γ is given as

$$\frac{dv(\gamma)}{d\gamma} = \frac{d}{d\gamma}\int_0^\infty \frac{1}{2\gamma}\exp(-x^2)(\gamma^2 + x^2)dx =$$
$$\int_0^\infty \frac{d}{d\gamma}\Big[\frac{1}{2\gamma}\exp(-x^2)(\gamma^2 + x^2)\Big]dx =$$
$$\frac{1}{2}\int_0^\infty \exp(-x^2) - \frac{\exp(-x^2)x^2}{\gamma^2}dx =$$
$$\frac{1}{2}\int_0^\infty \exp(-x^2)dx - \frac{1}{2}\int_0^\infty \frac{\exp(-x^2)x^2}{\gamma^2}dx$$

Then setting the derivative of $v(\gamma)$ to zero to find potential extreme values we have

$$\frac{1}{2}\int_0^\infty \exp(-x^2)dx - \frac{1}{2}\int_0^\infty \frac{\exp(-x^2)x^2}{\gamma^2}dx = 0 \tag{7}$$

$$\underbrace{\int_0^\infty \exp(-x^2)dx}_{I_1} = \underbrace{\int_0^\infty \frac{\exp(-x^2)x^2}{\gamma^2}dx}_{I_2} \tag{8}$$

First we evaluate I_2

$$I_2 = \frac{1}{\gamma^2}\int_0^\infty \exp(-x^2)x^2 dx = \frac{1}{\gamma^2}\int_0^\infty \underbrace{x}_{g}\cdot\underbrace{\exp(-x^2)x}_{f'}\,dx =$$

$$\frac{1}{2\gamma^2}[\underbrace{-\exp(-x^2)}_{g}\underbrace{x}_{f}]_0^\infty - \frac{-1}{2\gamma^2}\int_0^\infty \underbrace{\exp(-x^2)}_{g}\cdot\underbrace{1}_{f'}dx = \frac{1}{2\gamma^2}I_1$$

Noting how the integrand $\exp(-x^2)$ is positive and nonzero for all values larger than 0, it's evident how I_1 will not equal zero. Meaning that the expression in eq. 8 can be rewritten as

$$I_1 = \frac{1}{2\gamma^2}I_1 \rightarrow \gamma = \frac{1}{\sqrt{2}}, \qquad \text{since } \gamma > 0$$

Looking at the second derivative of $v(\gamma)$ we have

$$\frac{d^2v(\gamma)}{d\gamma^2} = \frac{d^2}{d\gamma^2}\left[\frac{-1}{2}\int_0^\infty \frac{\exp(-x^2)x^2}{\gamma^2}dx\right] = \int_0^\infty \frac{\exp(-x^2)x^2}{\gamma^3}dx > 0 \tag{9}$$

Hence $\hat{\gamma} = \frac{1}{\sqrt{2}}$ is a minimum. Since $v(\gamma)$ is C^1 for all $\gamma > 0$ it must also be a global minimum, and thus the value we are looking for. Figure 1 illustrates how both the variance in $\hat{Z}$ is reduced and how the empirical mean approaches Z when increasing the number of particles.

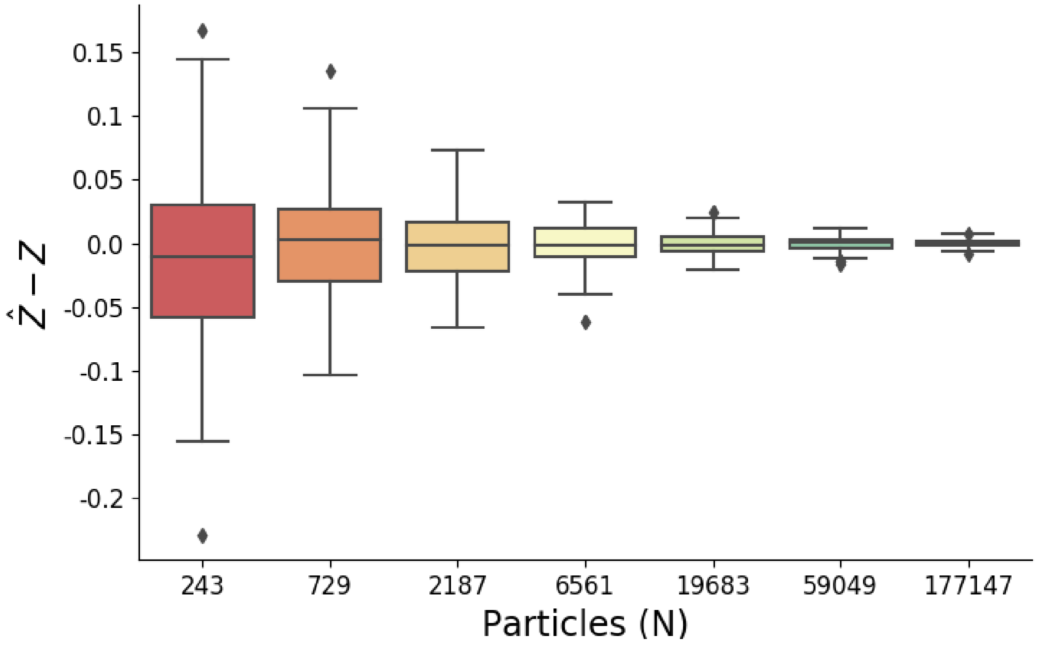
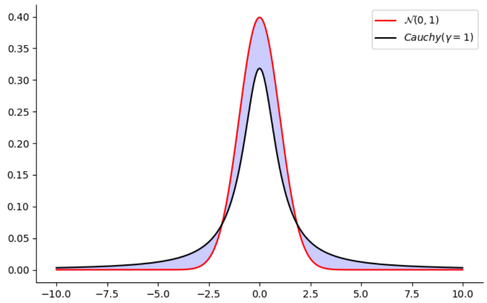
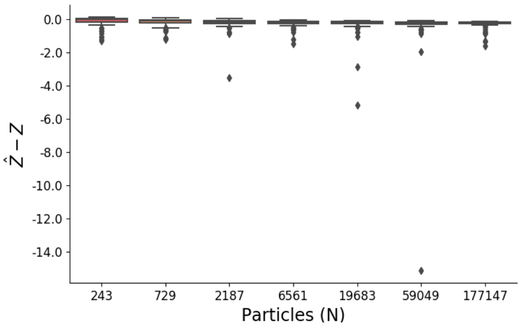


Figure 1: Boxplot of the difference between the true value of Z and the values estimated from Importance Sampling ($\hat{Z}$). Each box represents the usage of a varying number of particles (N), from which $\hat{Z}$ was computed.

c) **Given :** $\pi(x)$ is Cauchy (see **??** with $\gamma = 1$ and target ($q(x)$) is standard Gaussian.
Objective: Investigate how the standard Gaussian performs as a proposal distribution for the normalization constant of the Cauchy distribution.



(a) The Cauchy (black) and Normal (red) distribution



(b) Same type of figure as in Fig. 1 but with the standard Gaussian as proposal distribution and the Cauchy as target distribution.

Solution: Since the Gaussian have thinner tails than the Cauchy distribution, and approaches zero faster there are regions where the pdf of the proposal is smaller than that of the target distribution (see Fig. 2a)). Thus the quotient $\pi(x^i)/q(x^i)$ used during the importance sampling may obtain extremely high values, increasing the variance significantly. Such growth of the variance can be seen in Fig. **??**), where multiple outliers are observed.

2 H.2. Particle Filter for a linear Gaussian state-space model

a) **Given :** The state space model (*) as given in 10

$$(*)\begin{cases} X_t = & 0.8X_{t-1} + V_t, \quad V_t \sim \mathcal{N}(0, 0.5) \\ Y_t = & 2X_t + E_t, \quad\quad E_t \sim \mathcal{N}(0, 0.1) \end{cases} \tag{10}$$

Objective : Rewrite (*) as given in eq 2a-b) in the hand in assignment, and simulate data for $\mathcal{T} = \{t_i\}_0^{2000}$.
Solution : Eq. 10 can be rewritten as follows

$$X_t|X_{t-1} = x_{t-1} \sim \mathcal{N}(0.8x_{t-1}, 0.5) \tag{11}$$

$$Y_t|X_t = x_t \sim \mathcal{N}(2x_t, 0.01) \tag{12}$$

$$\tag{13}$$

Fig. **??** illustrates the result from simulating a trajectory $\mathcal{X}$, including both the hidden values (X) and observed values (Y) for all time points in $\mathcal{T}$.

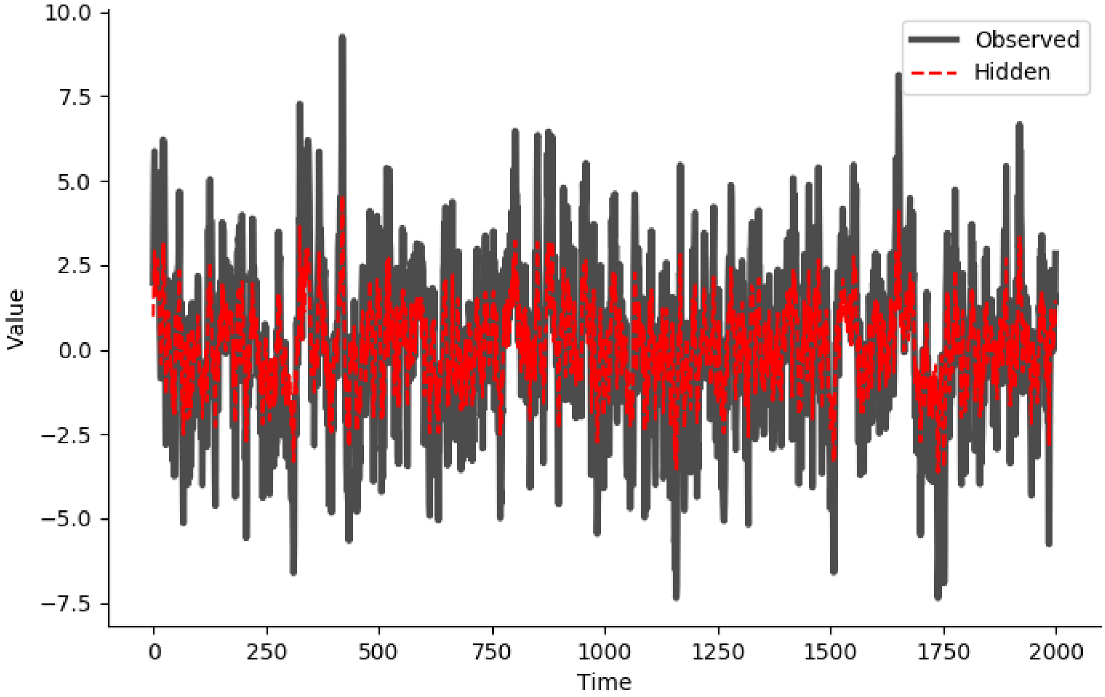


Figure 3: Simulated data for all time points $t \in \mathcal{T}$, with the hidden values (X) in red and observed values (Y) in black.

b) Comparing the analytical solution provided by the Kalman Filter (as given in eq 3-4) in the hand in assignment with the true 'hidden' values. It's evident how these overlap well for all time points in $\mathcal{T}$, with the difference at the different time points osciliatting around 0. This is further illustrated in Fig. 4

A particle filter would not be expected to be closer to the true trajectory, given how the Kalman Filter is designed to be a MMSE (minimum mean square error) estimator and from the inherent design of the filter it follows that the Kalman Filter is the optimal filter for such linear Gaussian problems.

c) Fig 5 shows the absolute difference between the mean (top row) and variance (bottom row) for each time step, using $N \in \{10, 50, 100, 2000, 5000\}$ particles in the bootstrap particle filter. What is evident is how the difference for both statistics (mean and variance) decreases as the number of particle grows, as expected by the design of the filter. Table 1 represent the same results as Fig. 5 but where the average of the absolute difference has been computed over all time points for the mean respectively variance. ngfi

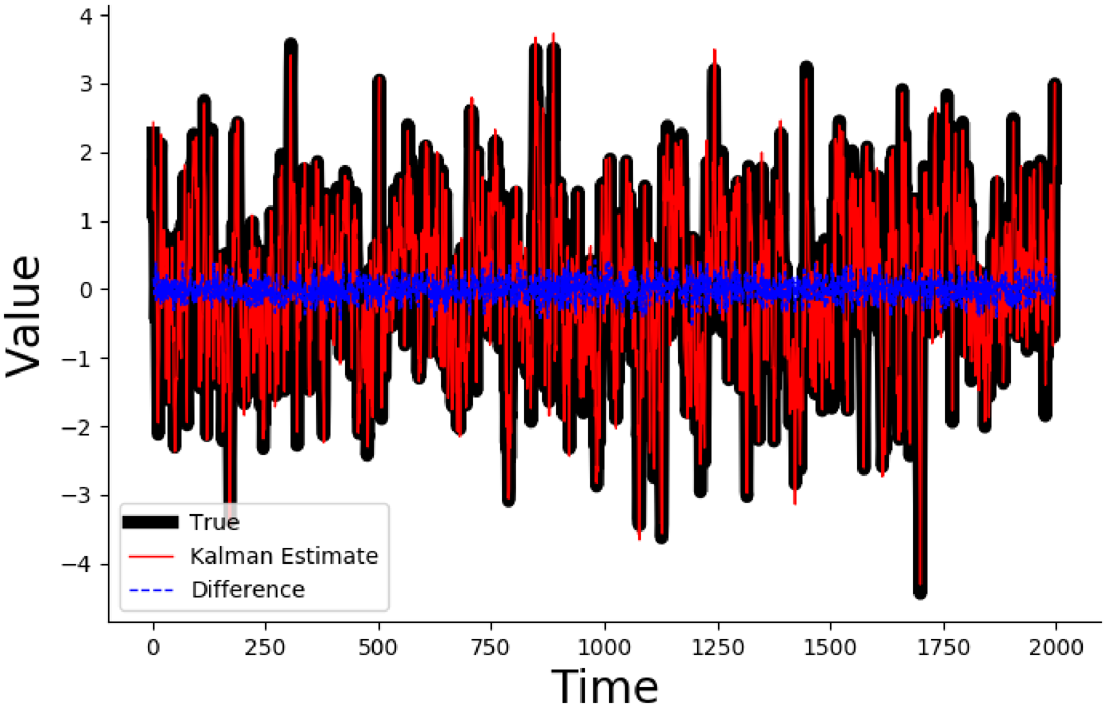


Figure 4: Trajectory generated by the Kalman Filter (red) compared to the analytical solution (black) with the difference between the two given in blue.

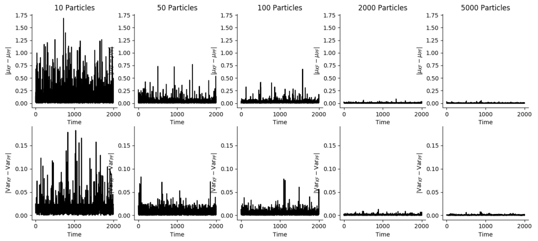


Figure 5: Absolute difference of Mean (top) and Variance (bottom) between BPF and Kalman Filter estiamtes of the hidden values.

N	Mean	Variance
10	0.153023	0.004702
50	0.037386	0.002453
100	0.020296	0.001920
2000	0.003657	0.001184
5000	0.002305	0.001116

Table 1: Average absolute difference to the Kalman Filter (KF) taken over all time points for the Bootstrap Particle Filter

d) **Objective :** Derive the local optimal proposals for the system in Eq. 10, implement the fully adapted particle filter and compare to the bootstrap particle filter.
Solution : For a linear gaussian state space model (LGSS), like that in Eq. 10 there exists an closed form of the local omptimal solution which will be derived (for the explicit system in this task) below. We will begin with the *propagation proposal* $q(x_t|y_t, x_{t-1}) = p(x_t|y_t, x_{t-1})$.

Due to the hierarchical relationship between y_t and x_t, i.e $y_t|x_t \sim \mathcal{N}(2x_t, 0.01)$ and x_t also being normal distributed, the two variables are jointly Gaussian when conditioned on x_{t-1}. Applying Theorem 9 from the course literature (Appendix Eq. B.52) we obtain the following expression the joint distribution together with the definitions given in 10.

$$p(x_t|y_t, x_{t-1}) \sim \mathcal{N}(\mu_{x_t|y_t}, \sigma^2_{x_t|y_t}) \tag{14}$$

$$\text{with} \tag{15}$$

$$\sigma^2_{x_t|y_t} = 0.5 - \frac{0.5^2 2^2}{0.1 + 2^2 \cdot 0.5} \tag{16}$$

$$\mu_{x_t|y_t} = 0.8x_{t-1} + \frac{0.5}{0.1 + 2^2 \cdot 0.5} \cdot 2 \cdot (y_t - 2 \cdot 0.8x_{t-1}) \tag{17}$$

$$\tag{18}$$

For the adjustment multipliers used to generate the resampling weights we have that

$$\nu^i_{t-1} = p(y_t|x^i_{t-1}) = \mathcal{N}(2 \cdot 0.8x_{t-1}, 2^2 \cdot 0.5 + 0.1) = \mathcal{N}(1.6x_{t-1}, 2.1) \tag{19}$$

Where we have used the following two identities

$$Z \sim \mathcal{N}(a\mu_x + b, a^2\sigma^2) \text{ if } Z = aX + b \text{ with } X \sim \mathcal{N}(\mu_x, \sigma^2_x) \tag{20}$$

$$Y \sim \mathcal{N}(\mu_z, \sigma^2_z + \sigma^2_y) \text{ if } Y \sim \mathcal{N}(Z, \sigma^2_y) \text{ with } Z \sim \mathcal{N}(\mu_z, \sigma^2_z) \tag{21}$$

The results from the implemented fully adapted particle filter (APF) are given in Table 2 whilst Fig. 6 compares these results to those obtained when using the bootstrap particle filter (BPF). As expected, we see how both the average mean and variance difference is lower in the fully adapted particles filter, with the effect being reduced the more particles that are introduced - meaning it's most prominent when employing few particles.

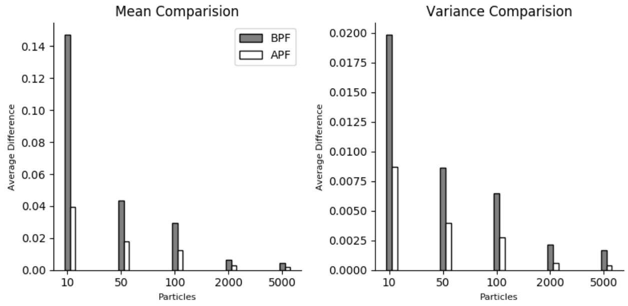


Figure 6: Average Mean (left) and Variance difference between the Fully Adapted Particle Filter (APF) and the Bootstrap Particle Filter (BPF)

	Mean	Variance
10	0.045969	0.018927
50	0.029285	0.021524
100	0.026824	0.022098
2000	0.023794	0.022341
5000	0.023466	0.022403

Table 2: Average absolute difference to the Kalman Filter (KF) taken over all time points for the Bootstrap Particle Filter

e) **Objective :** Visualize the genealogy of 100 particles when using the fully adapted particles filter. **Results :** Fig. 7 depicts all 100 trajectories of the particles used in the fully adapted particle filter. Only the first 250 time points are shown, but the genealogy of the particles found at $T = 2000$ is shown in red. The middle picture zooms in on the first 5 time points, whilst the rightmost picture looks at the time points in the interval $245 - 250$.

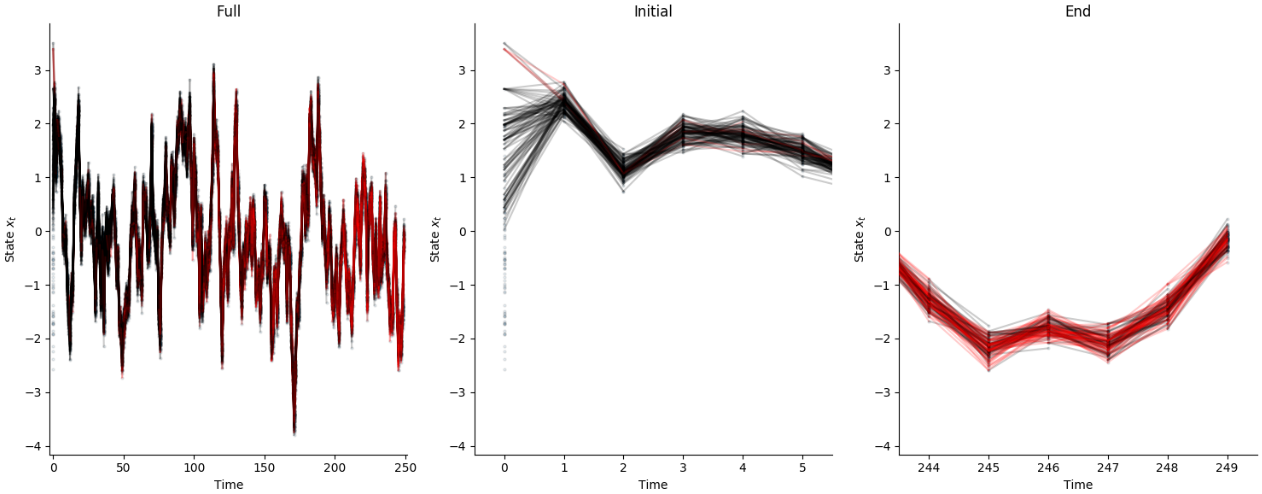


Figure 7: Trajectories generated when using the fully adapted particle filter and *Multinomial Resampling*, visualized from time point 0 to 250. Red trajectories belong to those particles that are found at the last time point $T = 2000$.

f) **Objective :** Use systematic resampling rather than multinomial sampling and plot the trajectories.
Results : Systematic resampling was implemented in accordance with algorithm **??**

Algorithm 1: Systematic Resampling

```
Draw $u \sim \mathcal{U}(0,1)$
for $i \in \{1..N\}$ do
    $u[i] = \frac{i-1+u}{N}$
    Find $r$ s.t. $u[i] \in [\sum_k^{r-1} w_k, \sum_k^r w_k)$
    $a[i] \leftarrow r$, with $r$ s.t.
end
```

Just as for the multinomial sampling, the trajectories can be visualized (see Fig **??**), where we can see how not only one particle is the primary ancestor to those particles present at the final time point - but rather these come from multiple origins.

Let the term *primary ancestor index* be defined as the index of the particle at the initial time point $t = 0$ for a particle observed at time point t.

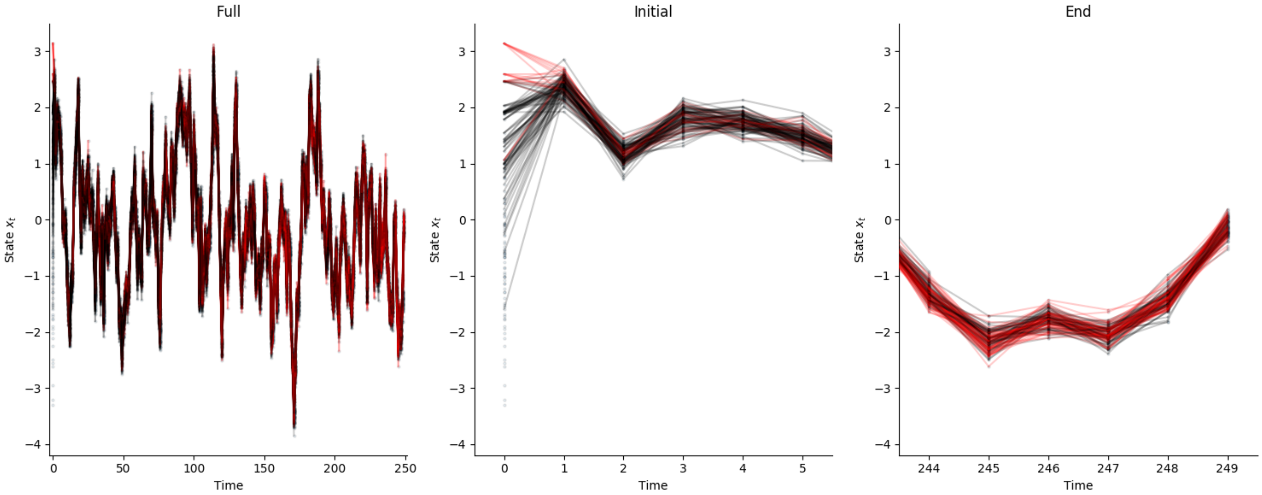


Figure 8: Trajectories generated when using the fully adapted particle filter and *Systematic Resampling*, visualized from time point 0 to 250. Red trajectories belong to those particles that are found at the last time point $T = 2000$.

By studying how many unique primary ancestor indices that are present at each time point, we can compare the path degeneracy between the two methods. Fig. 9 provides a visualization of the previously described comparison, where the number of unique primary ancestors of the particles at time point t are given on the y-axis, whilst time points are given on the x-axis. Systematic resampling outperforms multinomial resampling, with the latter having only one unique primary ancestor index around $t = 250$ whilst the former maintains multiple primary ancestors throughout the whole trajectory whilst also exhibiting a slower emergence of degeneracy.

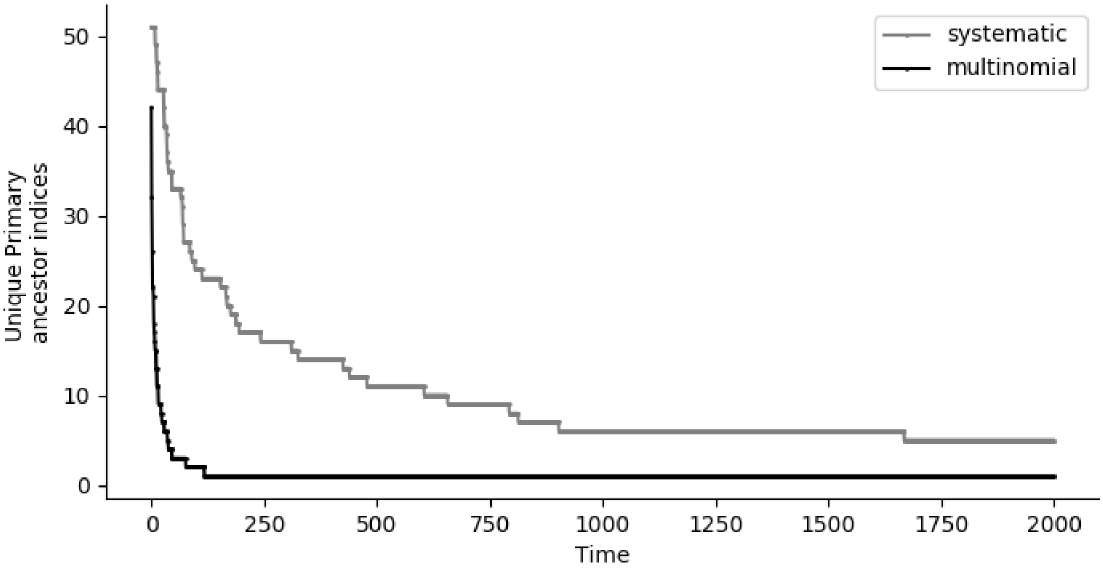
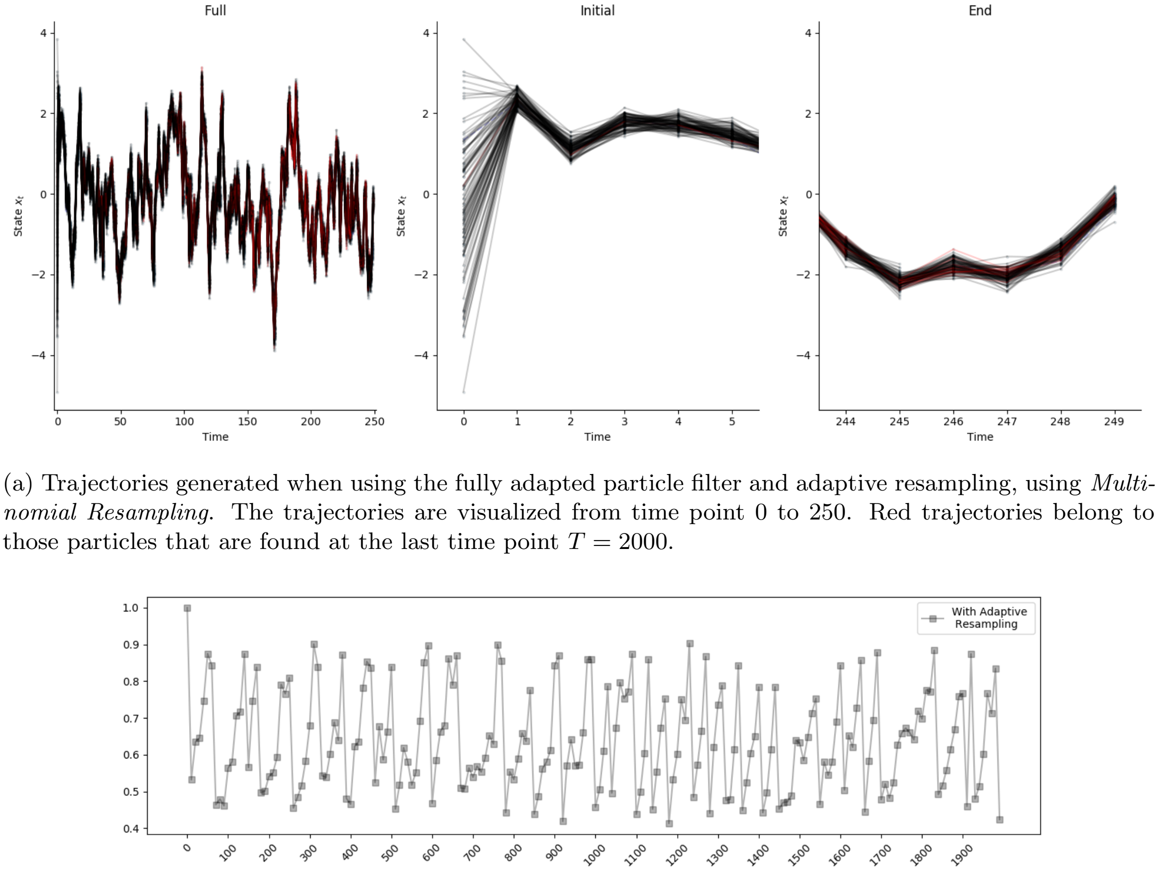


Figure 9: Number of unique primary ancestor indices at each time point when using Systematic (gray) respectively Multinomial (black) resampling.

Such behaviour is expected since the systematic resampling algorithm is designed to reduce the variance in the resampling, by imposing certain restrictions on the set of ancestor indices that can be sampled for offspring.

g) **Objective :** Use adaptive resampling with a threshold of $N_{thrs} = N/2 = 50$ and study the particle geneaology together with N_{ESS}/N as a function of time.
Results: Implementing the adaptive resampling scheme the genealogy presented in Fig. 10a whilst the quotient between the effective sample size and the number of particles present are visualized in 10b, where it a resampling event is characterized by a sharp peak in the graph.



(a) Trajectories generated when using the fully adapted particle filter and adaptive resampling, using *Multinomial Resampling*. The trajectories are visualized from time point 0 to 250. Red trajectories belong to those particles that are found at the last time point $T = 2000$.

(b) Quotient of N_{ESS}/N as a function of time

3 Parameter estimation in the stochastic volatility model

a) **Objective :** Given the stochastic volatility model (See eq. 23)

$$X_t|(X_t = x_{t-1} \sim \mathcal{N}(x_t|\phi x_{t-1}, \sigma^2) \tag{22}$$

$$Y_t|(X_t = x_t) \sim \mathcal{N}(y_t|0, \beta^2 \exp(x_t)) \tag{23}$$

Use a bootstrap particle filter to estimate the log-likelihood for values of $\phi \in [0, 1]$, given that $\sigma = 0.16$ and $\beta = 0.7$. Compute 10 estimates for each grid point and visualize the result using a boxplot.

Solution : Since the log-likelihood for $y_{1:T}$ can be derived according to the following procedure :

$$p(y_{1:T}|\theta) = \prod_1^T p(y_t|y_{1:t-1}, \theta) = \prod_1^T \int \underbrace{p(y_t|x_t, \theta)}_{\text{M.P.}} p(x_t|y_{1:t-1}, \theta) dx_t \tag{24}$$

Using the Bootstrap particle filter as an approximation of the conditional probability

$$p(x_t|y_{1:t-1}, \theta) \approx \frac{1}{N} \sum_{i=1}^N \delta_{x_t^i}(x_t) \tag{25}$$

Then inserting it into the integral in eq 24, allows us to approximate the likelihood

$$p(y_{1:T}) \approx \hat{p}(y_{1:T}) = \prod_1^T \frac{1}{N} \sum_{i=1}^N p(y_t|x_t^i, \theta) = \prod_1^T \frac{1}{N} \sum_{i=1}^N \hat{w}_t^i \tag{26}$$

And finally taking the logarithm of the likelihood, gives us the expression

$$\log \hat{p}(y_{1:T}) = \sum_{t-1}^T \left(\log \sum_{i=1}^N \hat{w}_t^i - \log N \right) \tag{27}$$

With $\hat{p}$ being the probability estimated using the bootstrap particle filter, and $\hat{w}_t^i$ representing the *unnormalized* weights at time t. Using this expression allows us to estimate the log likelihood for the parameter ϕ

Results : The log-likelihood was estimated 10 times for each value in the set $\{x : x = 0.1 \cdot i, i \in N \cap [1, 10]\}$ (i.e. ϕ was equated to each of these values and the log-likelihood was estimated 10-times). The boxplot in Fig. 11 illustrates the results. Given how the likelihood is highest for $\phi = 1.0$, a value near this seems to be a good candidate for the true value of ϕ.

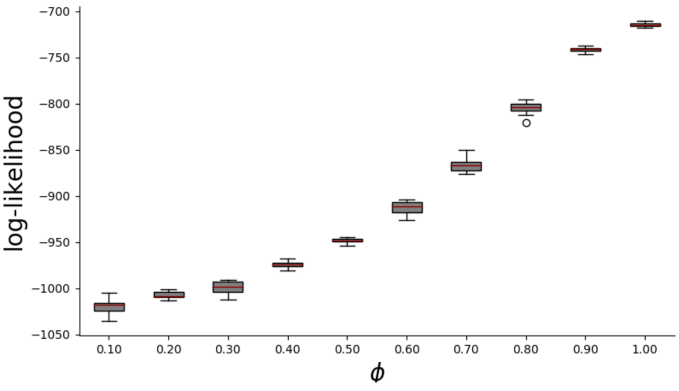


Figure 11: Boxplot of the log-likelihood estimates when using different values of ϕ and performing 10 iterations for each value.

b) **Objective :** Assuming that ϕ is known with a value set to 0.985 but that σ and β are unknown - Implement a PMH (Particle Metropolis Hastings) algorithm to compute the posterior of the two unknown

parameters and visualize the result in histograms.
Given : Both unknown parameters are assumed to have inverse gamma priors as defined in 30.

$$\sigma^2 \sim \mathcal{IG}(a = 0.01, b = 0.01) \tag{28}$$

$$\beta^2 \sim \mathcal{IG}(a = 0.01, b = 0.01) \tag{29}$$

$$\mathcal{IG}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(\frac{-b}{x}) \tag{30}$$

Solution : The implementation of the PMH algorithm is simple and can be summarized as in Algorithm 2

Algorithm 2: PMH Algorithm

```
Set initial conditions:
$\sigma^2[0] \sim \mathcal{U}(0,1)$
$\beta^2[0] \sim \mathcal{U}(0,1)$
Compute $L[0] = p(y_{1:T}|\sigma^2[0], \beta^2[0])$, (using the BPF)
for $i \in \{1..N\}$ do
    Sample proposal $\hat{\sigma}^2 \sim \mathcal{N}(\sigma^2[i-1], 1)$
    Sample proposal $\hat{\beta}^2 \sim \mathcal{N}(\beta^2[i-1], 1)$
    Compute $\hat{L} = p(y_{1:T}|\hat{\sigma}^2, \hat{\beta}^2)$ (using the BPF)
    Let $\alpha = \min\left\{1, \frac{\hat{L}p(\hat{\sigma}^2)p(\hat{\beta}^2)}{L[i-1]p(\sigma^2[i-1])p(\beta^2[i-1])}\right\}$
    Sample $u \sim (0,1)$
    if $\alpha \geq u$ and $\sigma^2, \beta^2 > 0$ then
        $\sigma^2[i] \hookleftarrow \hat{\sigma}^2$
        $\beta^2[i] \hookleftarrow \hat{\beta}^2$
        $L[i] \hookleftarrow \hat{L}$
    else
        $\sigma^2[i] \hookleftarrow \sigma^2[i-1]$
        $\beta^2[i] \hookleftarrow \beta^2[i-1]$
        $L[i] \hookleftarrow L[i-1]$
    end
end
```

Results : A total of 1000 particles were used in the PMH, where the proposal distribution was set to a Gaussian Random walk (as requested) with a step size of 0.1^2. A total 2500 iterations were performed where the first 40% were discarded as a "burn-in" period. Initial parameters were sampled from two uniform distributions; $\sigma^2[0] \sim \mathcal{U}(0, 0.3)$ and $\beta^2[0] \sim \mathcal{U}(0, 1)$. Fig. 12 illustrates the marginal distributions of the two parameters, where the mean is indicated by a red line, located at 0.01855 for σ^2 and 0.55038 for β^2

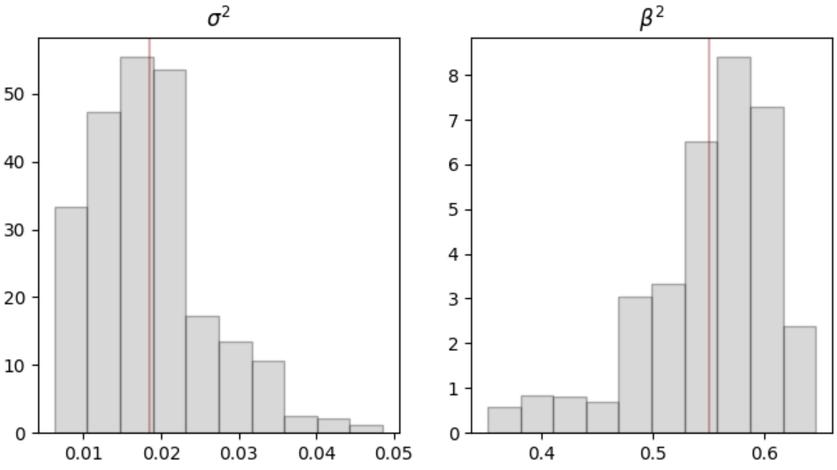


Figure 12: Marginal distributions of σ^2 and β^2 with the means being indicated in red.

c) **Objective :** Implement the Particle Gibbs sampler to sample from the posterior distribution of $p(\sigma^2, \beta^2|y_{1:T})$.

Results : Upon implementing the Particle Gibbs Sampler the conjugate distributions for σ^2 and β^2, are taken as those given in Eq. 8a-b in the Hand-In Instructions, a bootstrap particle filter with multinomial resampling in the particle Gibbs kernel. The bootstrap particle filter is modified as to allowed conditioning on a specific trajectory, which is "kept alive" as the particles are propagated in the filtering procedure. A total of 200 particles were used and 5000 iterations were performed, where the first 40% were discarded as a burn-in period. Fig. 13 show the distribution of σ^2 and β^2-values as sampled from the posterior.

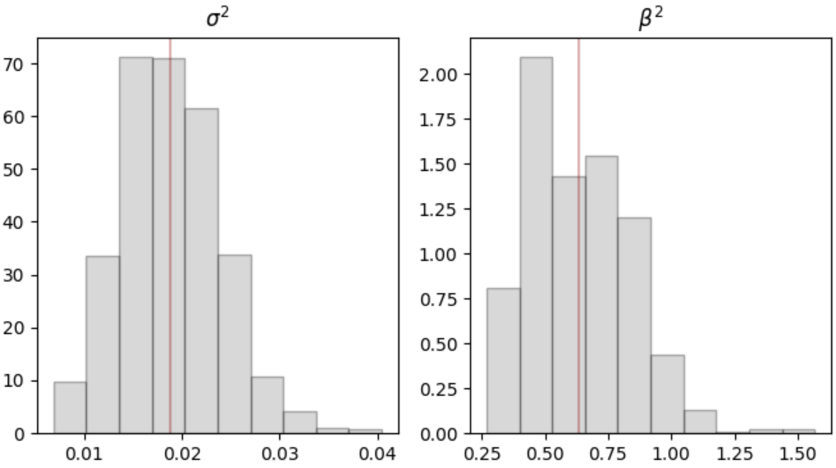


Figure 13: Posterior distribution of σ^2 and β^2. Red lines indicate the mean value (0.0192 for σ^2 and 0.673 for β^2

4 SMC Sampler

a) **Objective :** To implement a SMC sampler to estimate the normalizing constant of $\pi(x)$ which is proportional to $\gamma(\boldsymbol{x})$ (see Eq. 31.)

$$\gamma(\boldsymbol{x}) = \mathbb{I}((x_1, x_2) \in \mathcal{D}) \cos^2(x_1\pi) \sin^6(3\pi x_2) \exp(-30(x_1^2 + x_2^2)), \quad \mathcal{D} = [0,1] \times [0,1] \tag{31}$$

100 particles should be used, an effective sample size (ESS) of $0.7N = 70$ and an annealing sequence constructed such that resampling occurs at the magnitude of approximately every 10:th step. As an MCMC-kernel a Metropolis-Hastings scheme using a Gaussian Random walk should be used.
Results : For simplicity the annealing sequence $\{\gamma_k(\boldsymbol{x})\}_{k=0}^{K}$ was defined as

$$\gamma_k(x) = \gamma(x)^{k/K} \tag{32}$$

With $k = 0$ the initial distribution is reduced to $\gamma_0(\boldsymbol{x}) = 1$, and hence the normalizing constant Z_0 is given as

$$Z_0 = \int_{\mathcal{D}} \gamma_0(\boldsymbol{x}) d\boldsymbol{x} = 1 \tag{33}$$

Meaning that sampling from the initial distribution is equivalent to sampling from a uniform distribution defined over the unit square. Using a total of 100 particles, the recommended variance (0.02^2) was used in both directions, and K was set to 100. A step size of . Fig. 14 illustrates the distribution at 4 time points overlayed on the un-normalized density $\gamma(\boldsymbol{x})$.

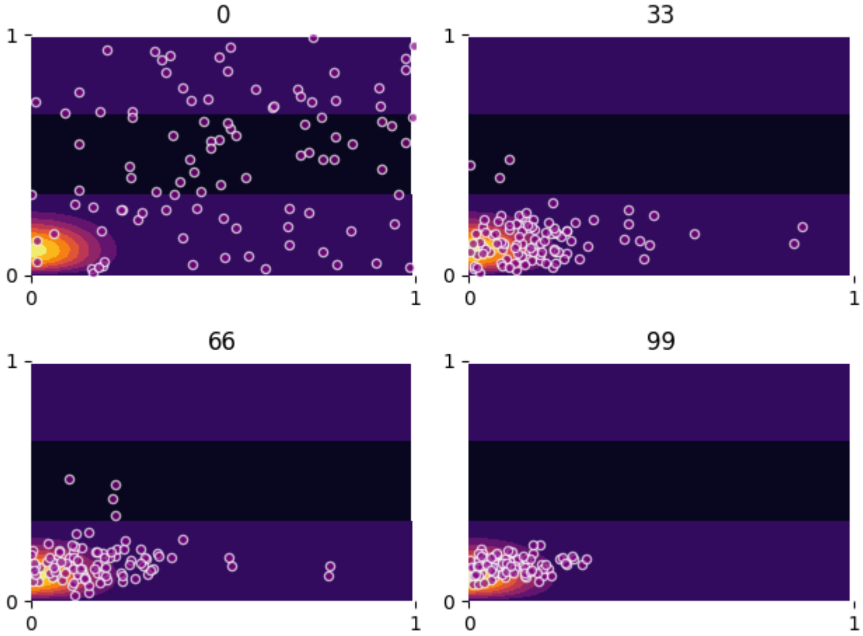


Figure 14: Position of particles at 4 different time-points, overlayed on the un-normalized density.

Plotting the ESS as a function of the iteration k results in the graph found in 15, where the resampling events are clearly visible as sharp peaks.

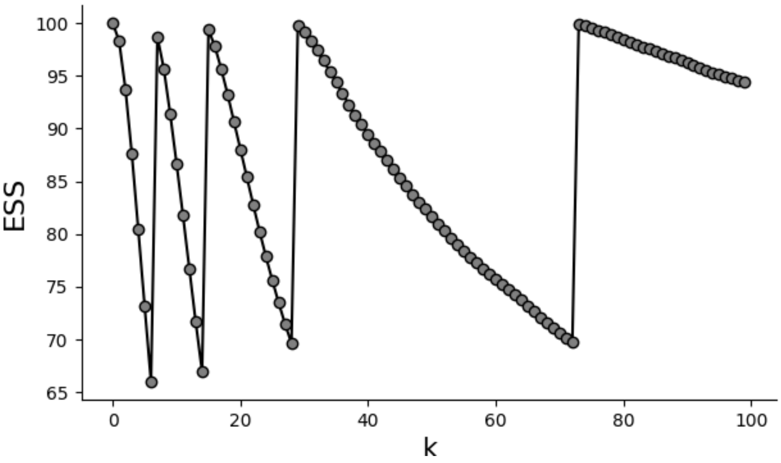


Figure 15: ESS as a function of iteration value k

With Z being defined as

$$Z = \int_{\mathcal{D}} \gamma(\boldsymbol{x}) d\boldsymbol{x} \tag{34}$$

Which can be approximated by

$$Z \approx \hat{Z}_K / Z_0 = \prod_{k=0}^{K} \frac{\hat{Z}_{k-1}}{Z_k}, \quad \frac{\hat{Z}_{k-1}}{Z_k} = \sum_{i=1} w_{k-1}^{i} \frac{\gamma_k x_{k-1}^{i}}{\gamma_{k-1}(x_{k-1}^{i})} \tag{35}$$

Z was hence estimated as 0.0066 .