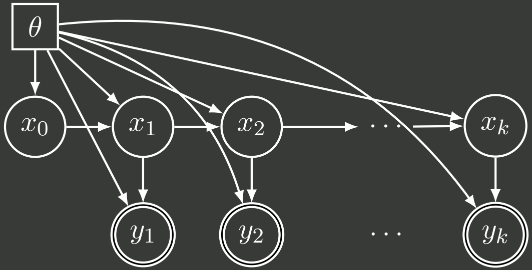
Advanced Computational Methods in Statistics
Lecture 5

O. Deniz Akyildiz

LTCC Advanced Course

December 11, 2023

Imperial College London



We are given the model

$$x_0 \sim \mu_\theta(x_0),$$
$$x_t | x_{t-1} \sim \tau_\theta(x_t | x_{t-1}),$$
$$y_t | x_t \sim g_\theta(y_t | x_t).$$

We looked at estimating θ given $y_{1:T}$.

- We have seen maximum likelihood approaches in the last session that would solve

$$\theta^\star \in \operatorname*{argmax}_{\theta \in \Theta} \log p(y_{1:T}|\theta).$$

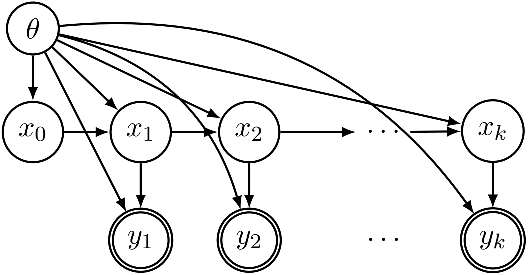
where

$$p(y_{1:T}|\theta) = \int p(y_{1:T}, x_{0:T}|\theta) \mathrm{d}x_{0:T}.$$

Today, we will first look at the Bayesian approach to this problem.

Problem definition

Recap – the model, the notation



We are given the model

$$\theta \sim p(\theta),$$
$$x_0 \sim \mu_\theta(x_0),$$
$$x_t|x_{t-1} \sim \tau(x_t|x_{t-1}, \theta),$$
$$y_t|x_t \sim g(y_t|x_t, \theta).$$

We aim at sampling from $p(\theta|y_{1:T})$.

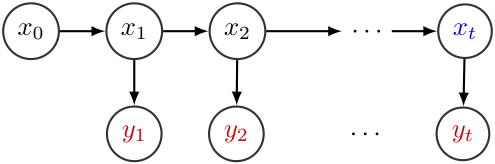
State-space models

Recall: Algorithmic principle

We are interested in estimating expectations,

$$(\varphi, \pi_t) = \int \varphi(x_t)\pi_t(x_t|y_{1:t})\mathrm{d}x_t = \int \varphi(x_t)\pi_t(\mathrm{d}x_t),$$

sequentially as new data arrives.



Algorithm:

Predict

$$\xi_t(\mathrm{d}x_t) = \int \pi_{t-1}(\mathrm{d}x_{t-1})\tau_t(\mathrm{d}x_t|x_{t-1})$$

Update

$$\pi_t(\mathrm{d}x_t) = \xi_t(\mathrm{d}x_t)\frac{g_t(y_t|x_t)}{p(y_t|y_{1:t-1})}.$$

A general algorithm to estimate expectations of any test function $\varphi(x_t)$ given $y_{1:t}$.

- Sampling: draw
$$\bar{x}_t^{(i)} \sim \tau_\theta(\mathrm{d}x_t | x_{t-1}^{(i)})$$
independently for every $i = 1, \ldots, N$.
- Weighting: compute
$$w_t^{(i)} = g_\theta(\bar{x}_t^{(i)}) / \bar{Z}_t^N$$
for every $i = 1, \ldots, N$, where $\bar{Z}_t^N = \sum_{i=1}^N g_\theta(\bar{x}_t^{(i)})$.
- Resampling: draw independently,
$$x_t^{(i)} \sim \tilde{\pi}_t(\mathrm{d}x) := \sum_i w_t^{(i)} \delta_{\bar{x}_t^{(i)}}(\mathrm{d}x) \quad \text{for } i = 1, ..., N.$$

$$\pi_{t-1}^N \underbrace{\rightarrow}_{\text{sampling}} \xi_t^N \underbrace{\rightarrow}_{\text{weighting}} \tilde{\pi}_t^N \underbrace{\rightarrow}_{\text{resampling}} \pi_t^N.$$

Bootstrap particle filter

Marginal likelihoods

Another quantity BPF can estimate is the marginal likelihood:

$$p(y_{1:t}|\theta) = \int p(y_{1:t}, x_{0:t}|\theta)\mathrm{d}x_{0:t}.$$

This quantity is useful for model selection and model comparison.

Recall that we have tbe factorisation:

$$p(y_{1:t}|\theta) = \prod_{k=1}^{t} p(y_k|y_{1:k-1}, \theta).$$

where

$$p(y_t|y_{1:t-1}, \theta) = \int g(y_t|x_t, \theta)\xi(x_t|y_{1:t-1}, \theta)\mathrm{d}x_t.$$

Recall that we can obtain the approximation of $\xi(x_t|y_{1:t-1}, \theta)$ by the particle filter using predictive particles $\bar{x}_t^{(i)} \sim \tau(x_t|x_{t-1}^{(i)}, \theta)$ as

$$p^N(\mathrm{d}x_t|y_{1:t-1}, \theta) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{x}_t^{(i)}}(\mathrm{d}x_t).$$

Therefore, given

$$p_\theta^N(\mathrm{d}x_t|y_{1:t-1},\theta) = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{x}_t^{(i)}}(\mathrm{d}x_t),$$

we get

$$p^N(y_t|y_{1:t-1},\theta) = \frac{1}{N}\sum_{i=1}^N g(y_t|\bar{x}_t^{(i)},\theta).$$

As a result, we can approximate

$$p^N(y_{1:t}|\theta) = \prod_{k=1}^t p^N(y_k|y_{1:k-1},\theta).$$

Remarkably, this estimate is unbiased:

$$\mathbb{E}[p^N(y_{1:t}|\theta)] = p(y_{1:t}|\theta),$$

for every fixed θ.

Parameter inference

A basic approach based on Metropolis-Hastings

Let us assume that we would like to sample from $p(\theta|y_{1:t})$

Parameter inference

A basic approach based on Metropolis-Hastings

Let us assume that we would like to sample from $p(\theta|y_{1:t})$

- We would normally use the factorisation

$$p(\theta|y_{1:t}) \propto p(y_{1:t}|\theta)p(\theta).$$

Parameter inference

A basic approach based on Metropolis-Hastings

Let us assume that we would like to sample from $p(\theta|y_{1:t})$

- We would normally use the factorisation

$$p(\theta|y_{1:t}) \propto p(y_{1:t}|\theta)p(\theta).$$

- Based on this, we could design a Metropolis-Hastings algorithm (with any proposal).

Parameter inference

A basic approach based on Metropolis-Hastings

Let us assume that we would like to sample from $p(\theta|y_{1:t})$

- We would normally use the factorisation

$$p(\theta|y_{1:t}) \propto p(y_{1:t}|\theta)p(\theta).$$

- Based on this, we could design a Metropolis-Hastings algorithm (with any proposal).

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

- Accept θ' with probability $\min\{1, r(\theta^{(i)}, \theta')\}$ and set $\theta^{(i+1)} = \theta'$.

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

- Accept θ' with probability $\min\{1, r(\theta^{(i)}, \theta')\}$ and set $\theta^{(i+1)} = \theta'$.
- Otherwise, set $\theta^{(i+1)} = \theta^{(i)}$.

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

- Accept θ' with probability $\min\{1, r(\theta^{(i)}, \theta')\}$ and set $\theta^{(i+1)} = \theta'$.
- Otherwise, set $\theta^{(i+1)} = \theta^{(i)}$.

Can this be applicable for state-space models?

The issue:

- We do not know $p(y_{1:t}|\theta)$ as this is an integral over $x_{0:t}$:

$$p(y_{1:t}|\theta) = \int p(y_{1:t}, x_{0:t}|\theta) \mathrm{d}x_{0:t}.$$

Parameter inference

A basic approach based on Metropolis-Hastings

The issue:

- We do not know $p(y_{1:t}|\theta)$ as this is an integral over $x_{0:t}$:

$$p(y_{1:t}|\theta) = \int p(y_{1:t}, x_{0:t}|\theta) \mathrm{d}x_{0:t}.$$

- We can approximate this integral using the particle filter:

$$p^N(y_{1:t}|\theta) = \frac{1}{N} \sum_{i=1}^{N} g(y_t|\bar{x}_t^{(i)}, \theta).$$

Parameter inference

A basic approach based on Metropolis-Hastings

The issue:

- We do not know $p(y_{1:t}|\theta)$ as this is an integral over $x_{0:t}$:

$$p(y_{1:t}|\theta) = \int p(y_{1:t}, x_{0:t}|\theta) \mathrm{d}x_{0:t}.$$

- We can approximate this integral using the particle filter:

$$p^N(y_{1:t}|\theta) = \frac{1}{N} \sum_{i=1}^{N} g(y_t|\bar{x}_t^{(i)}, \theta).$$

- Remarkably, plugging in unbiased estimates in Metropolis-Hastings ratios preserves the stationary measure (Andrieu et al., 2010).

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p^N(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p^N(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p^N(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p^N(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

- Accept θ' with probability $\min\{1, r(\theta^{(i)}, \theta')\}$ and set $\theta^{(i+1)} = \theta'$.

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p^N(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p^N(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

- Accept θ' with probability $\min\{1, r(\theta^{(i)}, \theta')\}$ and set $\theta^{(i+1)} = \theta'$.
- Otherwise, set $\theta^{(i+1)} = \theta^{(i)}$.

Recall the Metropolis-Hastings algorithm for this case:

- Given $\theta^{(i)}$, sample $\theta' \sim q(\theta'|\theta^{(i)})$.
- Compute the acceptance ratio

$$r(\theta^{(i)}, \theta') = \frac{p^N(y_{1:t}|\theta')p(\theta')q(\theta^{(i)}|\theta')}{p^N(y_{1:t}|\theta^{(i)})p(\theta^{(i)})q(\theta'|\theta^{(i)})}.$$

- Accept θ' with probability $\min\{1, r(\theta^{(i)}, \theta')\}$ and set $\theta^{(i+1)} = \theta'$.
- Otherwise, set $\theta^{(i+1)} = \theta^{(i)}$.

This is called the particle Metropolis-Hastings algorithm.

A few drawbacks of this approach:

- The algorithm is not very efficient as it requires a large number of particles to obtain a good approximation of $p(y_{1:t}|\theta)$.

A few drawbacks of this approach:

- The algorithm is not very efficient as it requires a large number of particles to obtain a good approximation of $p(y_{1:t}|\theta)$.
- Also, for every parameter sample $\theta^{(i)}$, a fresh run of the particle filter is required.

A few drawbacks of this approach:

- The algorithm is not very efficient as it requires a large number of particles to obtain a good approximation of $p(y_{1:t}|\theta)$.
- Also, for every parameter sample $\theta^{(i)}$, a fresh run of the particle filter is required.

We will now look at a completely online approach.

Let us discuss a meta-sampler that can be used to sample from $p(\theta|y_{1:t})$. First, let us try to use a naive importance sampler to sample from $p(\theta|y_{1:t})$ (forget for now about latents $x_{1:t}$).

How to develop an importance sampler for evolving $p(\theta|y_{1:t})$?

Let us recall the recursions:

$$p(\theta|y_{1:t}) = \frac{p(y_t|\theta)p(\theta|y_{1:t-1})}{p(y_t|y_{1:t-1})}.$$

Let us recall the recursions:

$$p(\theta|y_{1:t}) = \frac{p(y_t|\theta)p(\theta|y_{1:t-1})}{p(y_t|y_{1:t-1})}.$$

With these recursions in mind, we can indeed naively try to develop an importance sampler.

Let us choose a proposal: $q(\theta)$ and then perform importance sampling:

- Sample $\theta^{(i)} \sim q(\theta)$ for $i = 1, \ldots, N$.

Let us choose a proposal: $q(\theta)$ and then perform importance sampling:

- Sample $\theta^{(i)} \sim q(\theta)$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = \frac{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})}{q(\theta^{(i)})}.$$

Let us choose a proposal: $q(\theta)$ and then perform importance sampling:

- Sample $\theta^{(i)} \sim q(\theta)$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = \frac{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})}{q(\theta^{(i)})}.$$

- Normalise the weights:

$$\mathsf{w}_t^{(i)} = \frac{\mathsf{W}_t^{(i)}}{\sum_{j=1}^N \mathsf{W}_t^{(j)}}.$$

Let us choose a proposal: $q(\theta)$ and then perform importance sampling:

- Sample $\theta^{(i)} \sim q(\theta)$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = \frac{p(y_{1:t}|\theta^{(i)})p(\theta^{(i)})}{q(\theta^{(i)})}.$$

- Normalise the weights:

$$\mathsf{w}_t^{(i)} = \frac{\mathsf{W}_t^{(i)}}{\sum_{j=1}^N \mathsf{W}_t^{(j)}}.$$

Can we get a sequential structure in weights as in the particle filter case?

We have

$$W_{0:t}(\theta) = \frac{p(y_{1:t}|\theta)p(\theta)}{q(\theta)}.$$

We have

$$W_{0:t}(\theta) = \frac{p(y_{1:t}|\theta)p(\theta)}{q(\theta)}.$$

Unlike the particle filter case, we do not have a sequential structure in the weights. One can try

$$W_{0:t}(\theta) = p(y_t|y_{1:t-1}, \theta)W_{0:t-1}(\theta).$$

This means that we have to unroll it back to time zero:

$$W_{0:t}(\theta) = p(y_t|y_{1:t-1}, \theta)p(y_{t-1}|y_{1:t-2}, \theta) \cdots \frac{p(\theta)}{q(\theta)}.$$

Given

$$W_{0:t}(\theta) = p(y_t|y_{1:t-1}, \theta)p(y_{t-1}|y_{1:t-2}, \theta) \cdots \frac{p(\theta)}{q(\theta)}.$$

the practical weight computation would be:

$$\mathsf{W}_0^{(i)} = \frac{p(\theta^{(i)})}{q(\theta^{(i)})},$$

and

$$\mathsf{W}_t^{(i)} = p(y_t|y_{1:t-1}, \theta^{(i)})\mathsf{W}_{t-1}^{(i)}.$$

This would cause multiple issues:

- The algorithm is essentially putting samples into the space and just recomputing weights.

This would cause multiple issues:

- The algorithm is essentially putting samples into the space and just recomputing weights.
  - Samples do not move!

This would cause multiple issues:

- The algorithm is essentially putting samples into the space and just recomputing weights.
  - Samples do not move!
- Even if we introduce resampling at every stage, then still have the same problem.

This would cause multiple issues:

- The algorithm is essentially putting samples into the space and just recomputing weights.
  - Samples do not move!
- Even if we introduce resampling at every stage, then still have the same problem.
  - Samples do not move + are resampled.

This would cause multiple issues:

- The algorithm is essentially putting samples into the space and just recomputing weights.
  - Samples do not move!
- Even if we introduce resampling at every stage, then still have the same problem.
  - Samples do not move + are resampled.
  - Only one sample will survive.

This would cause multiple issues:

- The algorithm is essentially putting samples into the space and just recomputing weights.
  - Samples do not move!
- Even if we introduce resampling at every stage, then still have the same problem.
  - Samples do not move + are resampled.
  - Only one sample will survive.
- We need to introduce a new mechanism to move the samples around.

We need a way to *shake* the particles, without introducing too much error.

- Use a jittering kernel (Crisan and Míguez, 2014):

$$\kappa(\mathrm{d}\theta|\theta') = (1 - \epsilon_N)\delta_{\theta'}(\mathrm{d}\theta) + \epsilon_N \tau(\mathrm{d}\theta|\theta'), \qquad (1)$$

  to sample new particles $\theta_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$.
- We usually choose $\epsilon_N \leq \frac{1}{\sqrt{N}}$.
- τ can be simple, i.e., multivariate Gaussian or multivariate t distribution.

The jittered sampler:

- Sample $\bar{\theta}_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$ for $i = 1, \ldots, N$.

The jittered sampler:

- Sample $\bar{\theta}_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = p(y_t|y_{1:t-1}, \bar{\theta}_t^{(i)}),$$

The jittered sampler:

- Sample $\bar{\theta}_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = p(y_t|y_{1:t-1}, \bar{\theta}_t^{(i)}),$$

- Normalise the weights:

$$\mathsf{w}_t^{(i)} = \frac{\mathsf{W}_t^{(i)}}{\sum_{j=1}^N \mathsf{W}_t^{(j)}}.$$

- Resample:

$$\theta_t^{(i)} \sim \sum_{j=1}^N \mathsf{w}_t^{(j)} \delta_{\bar{\theta}_t^{(j)}}(\mathrm{d}\theta).$$

As you could guess, "compute the importance weights" step should be done using a particle filter.

- Sample $\bar{\theta}_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$ for $i = 1, \dots, N$.

As you could guess, "compute the importance weights" step should be done using a particle filter.

- Sample $\bar{\theta}_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = p^M(y_t|y_{1:t-1}, \bar{\theta}_t^{(i)}),$$

  using a particle filter with M particles.

As you could guess, "compute the importance weights" step should be done using a particle filter.

- Sample $\bar{\theta}_t^{(i)} \sim \kappa(\cdot|\theta_{t-1}^{(i)})$ for $i = 1, \ldots, N$.
- Compute the importance weights:

$$\mathsf{W}_t^{(i)} = p^M(y_t|y_{1:t-1}, \bar{\theta}_t^{(i)}),$$

  using a particle filter with M particles.
- Normalise the weights:

$$\mathsf{w}_t^{(i)} = \frac{\mathsf{W}_t^{(i)}}{\sum_{j=1}^N \mathsf{W}_t^{(j)}}.$$

- Resample:

$$\theta_t^{(i)} \sim \sum_{j=1}^N \mathsf{w}_t^{(j)} \delta_{\bar{\theta}_t^{(j)}}(\mathrm{d}\theta).$$

This algorithm is purely online.

Both approaches (pMCMC and nested PF) rely on unbiased marginal likelihoods.

Both approaches (pMCMC and nested PF) rely on unbiased marginal likelihoods.

Therefore, the unbiasedness property of PFs are crucial.

So far, we have looked at plenty of algorithms, but little theory.

So far, we have looked at plenty of algorithms, but little theory.

We will now prove L_2 bounds for

So far, we have looked at plenty of algorithms, but little theory.

We will now prove L_2 bounds for

- Perfect Monte Carlo

So far, we have looked at plenty of algorithms, but little theory.

We will now prove L_2 bounds for

- Perfect Monte Carlo
- Importance sampling

So far, we have looked at plenty of algorithms, but little theory.

We will now prove L_2 bounds for

- Perfect Monte Carlo
- Importance sampling
- Particle filters.

So far, we have looked at plenty of algorithms, but little theory.

We will now prove L_2 bounds for

- Perfect Monte Carlo
- Importance sampling
- Particle filters.

Perfect Monte Carlo

Let us assume that we have samples $x^{(k)} \sim \pi$ and we build the estimator

$$(\varphi, \pi) \approx (\varphi, \pi^N) = \frac{1}{N} \sum_{k=1}^{N} \varphi(x^{(k)}).$$

Perfect Monte Carlo

Let us assume that we have samples $x^{(k)} \sim \pi$ and we build the estimator

$$(\varphi, \pi) \approx (\varphi, \pi^N) = \frac{1}{N} \sum_{k=1}^{N} \varphi(x^{(k)}).$$

Theorem 1 (Perfect Monte Carlo)

Let φ be a bounded function. Then, for any $N \geq 1$,

$$\|(\varphi, \pi) - (\varphi, \pi^N)\|_2 \leq \frac{2\|\varphi\|_\infty}{\sqrt{N}}.$$

Proof.

We first provide the proof for $p = 2$ for simplicity. We rewrite the L_2 norm using its definition as,

$$\left\|(\varphi, \pi) - (\varphi, \pi^N)\right\|_2 = \left\|(\varphi, \pi) - \frac{1}{N}\sum_{k=1}^{N} \varphi\left(x^{(k)}\right)\right\|_2$$

$$= \mathbb{E}\left[\left|(\varphi, \pi) - \frac{1}{N}\sum_{k=1}^{N} \varphi\left(x^{(k)}\right)\right|^2\right]^{1/2}.$$

Writing explicitly, we have,

$$\mathbb{E}\left[\left|(\varphi, \pi) - \frac{1}{N}\sum_{k=1}^{N} \varphi\left(x^{(k)}\right)\right|^2\right] = \frac{1}{N^2}\mathbb{E}\left[\left|\sum_{i=1}^{N}\left(\varphi(x^{(i)}) - (\varphi, \pi)\right)\right|^2\right].$$

(cont.)

We define $S^{(i)} = \varphi(x^{(i)}) - (\varphi, \pi)$ and note that $\mathbb{E}[S^{(i)}] = 0$ and $S^{(i)}$ are independent random variables. We therefore have,

$$\mathbb{E}\left[\left|(\varphi,\pi) - \frac{1}{N}\sum_{k=1}^{N}\varphi\left(x^{(k)}\right)\right|^2\right] = \frac{1}{N^2}\mathbb{E}\left[\left|\sum_{i=1}^{N} S^{(i)}\right|^2\right],$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left|S^{(i)}\right|^2\right] \leq \frac{N4\|\varphi\|_\infty^2}{N^2},$$

since $\left|S^{(i)}\right| = \left|\varphi(x^{(i)}) - (\varphi,\pi)\right| \leq 2\|\varphi\|_\infty$. Therefore, we have,

$$\left\|(\varphi,\pi) - (\varphi,\pi^N)\right\|_2 \leq \frac{2\|\varphi\|_\infty}{\sqrt{N}},$$

■

Let us assume that we have samples $x^{(k)} \sim \pi$ and we build the estimator

$$(\varphi, \pi) \approx (\varphi, \pi^N) = \frac{1}{N} \sum_{k=1}^{N} \varphi(x^{(k)}).$$

Perfect Monte Carlo

Imperial College London

Let us assume that we have samples $x^{(k)} \sim \pi$ and we build the estimator

$$(\varphi, \pi) \approx (\varphi, \pi^N) = \frac{1}{N} \sum_{k=1}^{N} \varphi(x^{(k)}).$$

Theorem 2 (Perfect Monte Carlo)

If $\mathit{var}_\pi(\varphi) < \infty$*, then for any* $N \geq 1$*,*

$$\|(\varphi, \pi) - (\varphi, \pi^N)\|_2 \leq \frac{\sqrt{\mathit{var}_\pi(\varphi)}}{\sqrt{N}}.$$

where

$$\mathit{var}_\pi(\varphi) = \int \varphi^2(x)\pi(\mathrm{d}x) - \left(\int \varphi(x)\pi(\mathrm{d}x)\right)^2.$$

Proof.

Since (φ, π^N) is unbiased, then MSE is equal to the variance of the estimator. We therefore have,

$$\begin{aligned}\mathbb{E}\left[\left((\varphi, \pi) - (\varphi, \pi^N)\right)^2\right] &= \mathsf{var}_\pi[(\varphi, \pi^N)], \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathsf{var}_\pi[\varphi(x^{(i)})], \\ &= \frac{1}{N}\mathsf{var}_\pi[\varphi(X)].\end{aligned}$$

■

Consider the self-normalising IS estimator for (φ, π):

$$(\varphi, \tilde{\pi}^N) = \sum_{i=1}^{N} \mathsf{w}^{(i)} \varphi(x^{(i)}),$$

where $\mathsf{w}^{(i)} = \mathsf{W}^{(i)} / \sum_{j=1}^{N} \mathsf{W}^{(j)}$ and $\mathsf{W}^{(i)} = \Pi(x^{(i)}) / q(x^{(i)})$.

Theorem 3

Let φ be a bounded function. Then, for any $N \geq 1$,

$$\|(\varphi, \pi) - (\varphi, \tilde{\pi}^N)\|_2 \leq \frac{2\|\varphi\|_\infty \sqrt{\rho}}{\sqrt{N}}.$$

where

$$\rho = \chi^2(\pi||q) + 1.$$

where

$$\chi^2(\pi||q) = \int \left(\frac{\pi(x)}{q(x)} - 1 \right)^2 q(x) \mathrm{d}x.$$

Suggests that the discrepancy between π and q controls the L_2 error.

Proof. We first note the following inequalities,

$$
\begin{aligned}
|(\varphi, \pi) - (\varphi, \tilde{\pi}^N)| &= \left| \frac{(\varphi W, q)}{(W, q)} - \frac{(\varphi W, q^N)}{(W, q^N)} \right| \\
&\leq \frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|}{|(W, q)|} + |(\varphi W, q^N)| \left| \frac{1}{(W, q)} - \frac{1}{(W, q^N)} \right| \\
&= \frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|}{|(W, q)|} + \|\varphi\|_\infty |(W, q^N)| \left| \frac{(W, q^N) - (W, q)}{(W, q)(W, q^N)} \right| \\
&= \frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|}{(W, q)} + \frac{\|\varphi\|_\infty |(W, q^N) - (W, q)|}{(W, q)}.
\end{aligned}
$$

We take squares of both sides and apply the inequality $(a+b)^2 \leq 2(a^2+b^2)$ to further bound the rhs,

$$\cdots \leq 2\frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|^2}{(W,q)^2} + 2\frac{\|\varphi\|_\infty^2 |(W, q^N) - (W,q)|^2}{(W,q)^2}$$

We can now take the expectation of both sides,

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\tilde{\pi}^N)\right)^2\right] \leq \frac{2\mathbb{E}\left[\left((\varphi W, q) - (\varphi W, q^N)\right)^2\right]}{(W,q)^2} + \frac{2\|\varphi\|_\infty^2 \mathbb{E}\left[\left((W, q^N) - (W,q)\right)^2\right]}{(W,q)^2}.$$

Note that, both terms in the right hand side are perfect Monte Carlo estimates of the integrals.

Bounding the MSE of these integrals yields

$$\begin{aligned}
\cdots &\leq \frac{2}{N}\frac{(\varphi^2 W^2, q) - (\varphi W, q)^2}{(W,q)^2} + \frac{2\|\varphi\|_\infty^2}{N}\frac{(W^2, q) - (W,q)^2}{(W,q)^2}, \\
&\leq \frac{2\|\varphi\|_\infty^2}{N}\frac{(W^2,q)}{(W,q)^2} + \frac{2\|\varphi\|_\infty^2}{N}\frac{(W^2, q) - (W,q)^2}{(W,q)^2}.
\end{aligned}$$

Therefore, we can straightforwardly write,

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\tilde{\pi}^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty^2}{(W,q)^2}\frac{(W^2,q)}{N}.$$

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\tilde{\pi}^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty^2}{(W,q)^2}\frac{(W^2,q)}{N}.$$

Now it remains to show the relation of the bound to χ^2 divergence. Note that,

$$\begin{aligned}
\frac{(W^2,q)}{(W,q)^2} &= \frac{\int \frac{\Pi^2(x)}{q^2(x)}q(x)\mathsf{d}x}{\left(\int \frac{\Pi(x)}{q(x)}q(x)\mathsf{d}x\right)^2} \\
&= \frac{Z^2\int \frac{\pi^2(x)}{q^2(x)}q(x)\mathsf{d}x}{Z^2\left(\int \pi\mathsf{d}x\right)^2} \\
&= \mathbb{E}_q\left[\frac{\pi^2(X)}{q^2(X)}\right] := \rho.
\end{aligned}$$

Note that ρ is not exactly χ^2 divergence, which is defined as $\rho - 1$. Plugging everything into our bound, we have the result,

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty^2\rho}{N}.$$

■

L_2 bound for the particle filter

Theorem 4

Let φ be a bounded function and π_t^N be particle filter approximations of π_t. Then, for any $N \geq 1$,

$$\|(\varphi, \pi_t) - (\varphi, \pi_t^N)\|_2 \leq \frac{c_t \|\varphi\|_\infty}{\sqrt{N}}.$$

where $c_t < \infty$ is a constant independent of N.

L_2 bound for the particle filter

Imperial College London

This is an induction based proof. At time $t = 0$, particle filter just samples from the prior of the model π_0 and by perfect Monte Carlo result, we readily have

$$\|(\varphi, \pi_0) - (\varphi, \pi_0^N)\|_2 \leq \frac{c_0 \|\varphi\|_\infty}{\sqrt{N}}.$$

where $c_0 = 2$. Therefore, as an induction hypothesis, we assume

$$\|(\varphi, \pi_{t-1}) - (\varphi, \pi_{t-1}^N)\|_2 \leq \frac{c_{t-1} \|\varphi\|_\infty}{\sqrt{N}}.$$

Particle filter takes three steps. We need to bound them separately.

Prediction/sampling step: Recall the predictive measure

$$\xi(\mathrm{d}x_t) = \int \tau(\mathrm{d}x_t|x_{t-1})\pi(\mathrm{d}x_{t-1}).$$

We need to next prove that the predictive approximation

$$\xi^N(\mathrm{d}x_t) = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{x}_t^{(i)}}(\mathrm{d}x_t),$$

where $\bar{x}_t^{(i)} \sim \tau(\mathrm{d}x_t|x_{t-1}^{(i)})$ satisfies the L_2 bound

$$\|(\varphi,\xi^N) - (\varphi,\xi)\|_2 \leq \frac{c_{1,t}\|\varphi\|_\infty}{\sqrt{N}}.$$

$$\begin{aligned}\|(\varphi, \xi^N) - (\varphi, \xi)\|_2 &= \left\|(\varphi, \xi_t^N) - (\varphi, \tau_t \pi_{t-1})\right\|_2 \\ &\leq \left\|(\varphi, \xi_t^N) - (\varphi, \tau_t \pi_{t-1}^N)\right\|_2 \\ &+ \left\|(\varphi, \tau_t \pi_{t-1}^N) - (\varphi, \tau_t \pi_{t-1})\right\|_2,\end{aligned}$$

where

$$(\varphi, \tau_t \pi_{t-1}^N) = \frac{1}{N} \sum_{i=1}^N (\varphi, \tau_t^{x_{t-1}^{(i)}}).$$

We have to now separately bound two terms.

For the first term, we introduce the σ-algebra generated by the random variables $x_{0:t}^{(i)}$ and $\bar{x}_{1:t}^{(i)}$, $i = 1, \ldots, N$, denoted $\mathcal{F}_t = \sigma(x_{0:t}^{(i)}, \bar{x}_{1:t}^{(i)}, i = 1, \ldots, N)$. Since π_{t-1}^N is measurable w.r.t. $\mathcal{F}_{t-1}$, we can write

$$\mathbb{E}[(\varphi, \xi_t^N)|\mathcal{F}_{t-1}] = \frac{1}{N}\sum_{i=1}^N (\varphi, \tau_t^{x_{t-1}^{(i)}}) = (\varphi, \tau_t \pi_{t-1}^N).$$

Next, we define the random variables $S_t^{(i)} = \varphi(\bar{x}_t^{(i)}) - (\varphi, \tau_t \pi_{t-1}^N)$ and note that, conditional on $\mathcal{F}_{t-1}$, $S_t^{(i)}$, $i = 1, \ldots, N$ are zero-mean and independent. Then, the approximation error of ξ_t^N can be written as,

$$\mathbb{E}[\left|(\varphi, \xi_t^N\ ) - (\varphi, \tau_t \pi_{t-1}^N)\right|^2 |\mathcal{F}_{t-1}] = \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N S_t^{(i)}\right|^2 \middle| \mathcal{F}_{t-1}\right].$$

Using the fact that $S_t^{(i)}$ are conditionally zero-mean and independent, we can write,

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N S_t^{(i)}\right|^2 \middle| \mathcal{F}_{t-1}\right] = \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^N \left|S_t^{(i)}\right|^2 \middle| \mathcal{F}_{t-1}\right],$$

Moreover, since $\left|S_t^{(i)}\right| = \left|\varphi(\bar{x}_t^{(i)}) - (\varphi, \tau_t \pi_{t-1}^N)\right| \leq 2\|\varphi\|_\infty$, we have,

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N S_t^{(i)}\right|^2 \middle| \mathcal{F}_{t-1}\right] \leq \frac{1}{N^2} N 4\|\varphi\|_\infty^2 = \frac{4\|\varphi\|_\infty^2}{N}.$$

If we take unconditional expectations on both sides of the equation above, then we arrive at

$$\|(\varphi, \xi_t^N) - (\varphi, \tau_t \pi_{t-1}^N)\|_2 \leq \frac{\tilde{c}_1 \|\varphi\|_\infty}{\sqrt{N}}, \qquad (2)$$

where $\tilde{c}_1 = 2$ is a constant independent of N.

To handle the second term, we define $(\bar{\varphi}, \pi_{t-1}) = (\varphi, \tau_t \pi_{t-1})$ where $\bar{\varphi} \in B(\mathsf{X})$ and given by,

$$\bar{\varphi}(x) = (\varphi, \tau_t^x).$$

We also write $(\bar{\varphi}, \pi_{t-1}^N) = (\varphi, \tau_t \pi_{t-1}^N)$. Since $\|\bar{\varphi}\|_\infty \leq \|\varphi\|_\infty$, the induction hypothesis leads,

$$\begin{aligned} \|(\varphi, \tau_t \pi_{t-1}^N) - (\varphi, \tau_t \pi_{t-1})\|_2 &= \|(\bar{\varphi}, \pi_{t-1}^N) - (\bar{\varphi}, \pi_{t-1})\|_2 \\ &\leq \frac{c_{t-1} \|\varphi\|_\infty}{\sqrt{N}}, \end{aligned} \tag{3}$$

where c_{t-1} is a constant independent of N. Combining (2) and (3) yields,

$$\left\|(\varphi, \xi_t^N) - (\varphi, \xi_t)\right\|_2 \leq \frac{c_{1,t} \|\varphi\|_\infty}{\sqrt{N}} \tag{4}$$

where $c_{1,t} = c_{t-1} + 2 < \infty$ is a constant independent of N.

Weighting step: Next, we aim at bounding $\|(\varphi, \pi_t) - (\varphi, \tilde{\pi}_t^N)\|_2$ using (4). We have the weighted random measure,

$$\tilde{\pi}_t^N = \sum_{i=1}^N w_t^{(i)} \delta_{\bar{x}_t^{(i)}} \quad \text{where} \quad w_t^{(i)} = \frac{g_t(\bar{x}_t^{(i)})}{\sum_{i=1}^N g_t(\bar{x}_t^{(i)})}.$$

The integrals computed with respect to the weighted measure $\tilde{\pi}_t^N$ takes the form,

$$(\varphi, \tilde{\pi}_t^N) = \frac{(\varphi g_t, \xi^N)}{(g_t, \xi_t^N)}. \tag{5}$$

On the other hand, using Bayes theorem, integrals with respect to the optimal filter can also be written in a similar form as,

$$(\varphi, \pi_t) = \frac{(\varphi g_t, \xi_t)}{(g_t, \xi_t)}. \tag{6}$$

Using a similar argument as in the proof of importance sampling

$$\left|(\varphi, \tilde{\pi}_t^N) - (\varphi, \pi_t)\right| \leq \frac{1}{(g_t, \xi_t)} \left(\|\varphi\|_\infty \left|(g_t, \xi_t) - (g_t, \xi_t^N)\right| + \left|(\varphi g_t, \xi_t) - (\varphi g_t, \xi_t^N)\right| \right), \tag{7}$$

where $(g_t, \xi_t) > 0$ by assumption. Using Minkowski's inequality, we can deduce from (7) that

$$\left\|(\varphi, \tilde{\pi}_t^N) - (\varphi, \pi_t)\right\|_2 \leq \frac{1}{(g_t, \xi_t)} \left(\|\varphi\|_\infty \left\|(g_t, \xi_t) - (g_t, \xi_t^N)\right\|_2 + \left\|(\varphi g_t, \xi_t) - (\varphi g_t, \xi_t^N)\right\|_2 \right). \tag{8}$$

Noting that we have $\|\varphi g_t\|_\infty \leq \|\varphi\|_\infty \|g_t\|_\infty$, (4) and (8) together yield,

$$\left\|(\varphi, \pi_t) - (\varphi, \tilde{\pi}_t^N)\right\|_2 \leq \frac{c_{2,t}\|\varphi\|_\infty}{\sqrt{N}}, \qquad (9)$$

where

$$c_{2,t,p} = \frac{2\|g_t\|_\infty c_{1,t}}{(g_t, \xi_t)} < \infty$$

is a finite constant independent of N.

Resampling step: Finally, since the random variables which are used to construct π_t^N are sampled i.i.d from $\tilde{\pi}_t^N$, the argument for the base case can also be applied here to yield,

$$\left\|(\varphi, \tilde{\pi}_t^N) - (\varphi, \pi_t^N)\right\|_2 \leq \frac{c_{3,t}\|\varphi\|_\infty}{\sqrt{N}}, \tag{10}$$

where $c_{3,t} < \infty$ is a constant independent of N. Combining bounds (9) and (10) to obtain the final result, with $c_t = c_{2,t} + c_{3,t} < \infty$, concludes the proof. □

Thanks!

References I

- Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). "Particle markov chain monte carlo methods". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.3, pp. 269–342.
- Crisan, Dan and Joaquín Míguez (2014). "Particle-kernel estimation of the filter density in state-space models". In: *Bernoulli* 20.4, pp. 1879–1929.