Advanced Simulation - Lecture 13

Rémi Bardenet

6 March 2014

Last 4 lectures: Sequential Monte Carlo methods

- Infer θ by incorporating data sequentially.
  - SMC sampler for static data
- Generalize to infer hidden states in an HMM.
- Use SMC as a proposal in MH.

- Let data be $y_{1:T} = y_1, \ldots, y_T$, and the posterior

$$\pi(\theta) = p(\theta|y_{1:T}) \propto p(y_{1:T}|\theta)p(\theta).$$

- What do you remember of IS?
- During this lecture, our running example will be probit classification

$$y|x, \beta \sim \text{Bernoulli}(\Phi(\beta x))$$

where Φ is the cdf of $\mathcal{N}(0, 1)$.

Sequential importance sampling

Define $\pi_0(\theta) = p(\theta)$ and

$$\pi_t(\theta) = p(\theta|y_{1:t}) \text{ for } 1 \leq t \leq T.$$

Assume you have $\theta_t^1, \ldots, \theta_t^N \sim \pi_t$ iid, then

$$\pi_{t+p} \quad \approx \quad \frac{1}{N} \sum_{i=1}^{N} \frac{\pi_{t+p}(\theta_t^i)}{\pi_t(\theta_t^i)} \delta_{\theta_t^i}$$

But

$$\frac{\pi_{t+p}(\theta_t^i)}{\pi_t(\theta_t^i)} \propto \frac{p(y_{1:t+p}|\theta_t^i)}{p(y_{1:t}|\theta_t^i)} = p(y_{t+1:t+p}|y_{1:t}, \theta_t^i) =: w_{t+p}^i$$

Finally,

$$\pi_{t+p} \approx \frac{1}{\sum_{i=1}^{N} w_{t+p}^i} \sum_{i=1}^{N} w_{t+p}^i \delta_{\theta_t^i}.$$

If data is assumed iid given θ, then w_{t+p}^i is simply $\prod_{i=1}^{p} p(y_{t+i}|\theta_t^i)$.

Have you seen this algorithm before?

?

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
- Return $\hat{\pi}_T = \frac{1}{\sum_{j=1}^N w_T^j} \sum_{i=1}^N w_T^i \delta_{\theta_T^i}$

It's OK but...

Have you seen this algorithm before?

Importance sampling

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
- Return $\hat{\pi}_T = \frac{1}{\sum_{j=1}^N w_T^j} \sum_{i=1}^N w_T^i \delta_{\theta_T^i}$

It's OK but...

- We are doing simple batch importance sampling!
- Most weights are going to zero as t grows:
  - useless storage
  - high variance estimators

Results on probit model

- $T = 100$ samples,
- $N = 100$ particles.

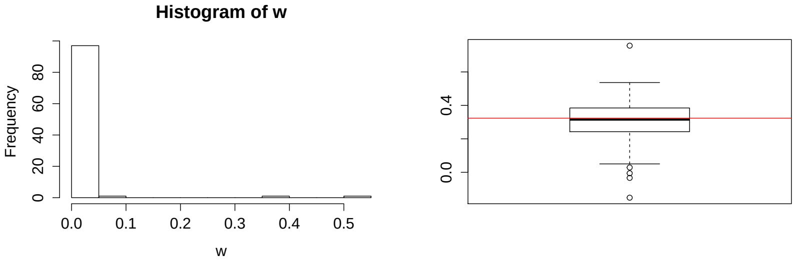

Figure: Left: example set of weights. Right: estimates of the posterior mean for 100 repetitions.

Systematic resampling

IS with systematic resampling

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - "Resample"

$$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$

    and reset $w_t^i = 1/N$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

It's OK but...

Systematic resampling

IS with systematic resampling

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - "Resample"
    $$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$
    and reset $w_t^i = 1/N$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

It's OK but...

- even higher variance as only a few particles will survive.
- If the support of the target moves with t, particles could even end up out of the support!

- $T = 100$ samples,
- $N = 100$ particles.

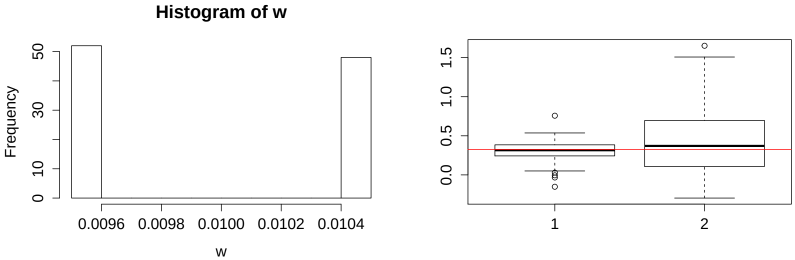

Figure: Left: example set of weights. Right: estimates of the posterior mean for 100 repetitions.

Resample only when needed

IS with occasional resampling

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - If most weights w_t^i are zero, then

$$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$

    and reset $w_t^i = 1/N$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

Resample only when needed

IS with occasional resampling

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - Compute $ESS_t = \left(\sum_{i=1}^N (w_t^i)^2\right)^{-1}$.
  - If $ESS_t < N/2$, then

$$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$

    and reset $w_t^i = 1/N$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

Resample only when needed

IS with occasional resampling

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{1:t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - Compute $ESS_t = \left(\sum_{i=1}^N (w_t^i)^2 \right)^{-1}$.
  - If $ESS_t < N/2$, then

$$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$

    and reset $w_t^i = 1/N$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

It's OK but...

- still potentially high variance as particles are highly correlated.

Results on probit model

- $T = 100$ samples,
- $N = 100$ particles.

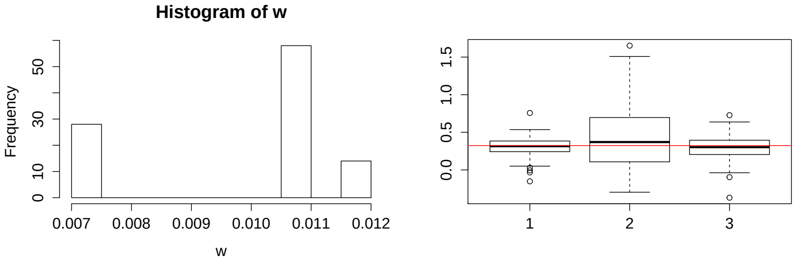

Figure: Left: example set of weights. Right: estimates of the posterior mean for 100 repetitions.

The SMC sampler (Chopin, 2002)

SMC sampler

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - Compute $ESS_t = \left(\sum_{i=1}^N (w_t^i)^2 \right)^{-1}$.
  - If $ESS_t < N/2$, then

$$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$

    - reset $w_t^i = 1/N$ for all i
    - Move $\theta_t^i \sim K_t(\theta_t^i, \cdot)$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

The SMC sampler (Chopin, 2002)

SMC sampler

- Sample $\theta_0^1, \ldots, \theta_0^N \sim \pi_0$ iid and set $w_0^i = 1/N$ for all i.
- For $1 \leq t \leq T$,
  - Set $w_t^i \propto w_{t-1}^i p(y_t | y_{t-1}, \theta_t^i)$.
  - Set $\theta_t^i = \theta_{t-1}^i$.
  - Compute $ESS_t = \left(\sum_{i=1}^N (w_t^i)^2 \right)^{-1}$.
  - If $ESS_t < N/2$, then

$$\theta_t^1, \ldots, \theta_t^N \sim \frac{1}{\sum_{i=1}^N w_t^i} \sum_{i=1}^N w_t^i \delta_{\theta_t^i} \text{ iid}$$

    - reset $w_t^i = 1/N$ for all i
    - Move $\theta_t^i \sim K_t(\theta_t^i, \cdot)$ for all i
- Return $\hat{\pi}_T = \frac{1}{\sum_{i=1}^N w_T^i} \sum_{i=1}^N w_T^i \theta_T^i$

It's finally OK!

- K_t is chosen to leave π_t invariant.
- the "move" step rejuvenates the set of particles.

Results on probit model

- $T = 100$ samples,
- $N = 100$ particles.
- Move kernel is MH with wide proposal.

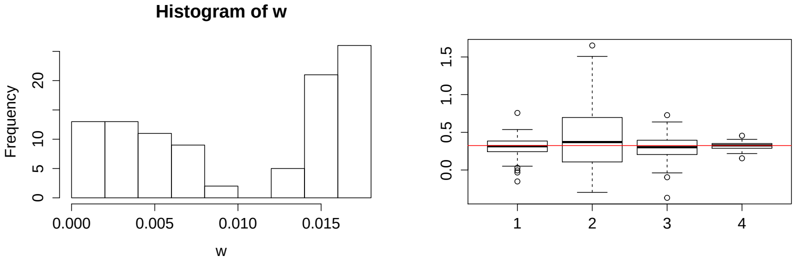

Figure: Left: example set of weights. Right: estimates of the posterior mean for 100 repetitions.

Teaser

- Next lecture, we will start generalizing this approach to inference in HMMs,

- Some keywords: high-dimensional inference for time-series data, Kalman filter, SMC filter, smoothing.