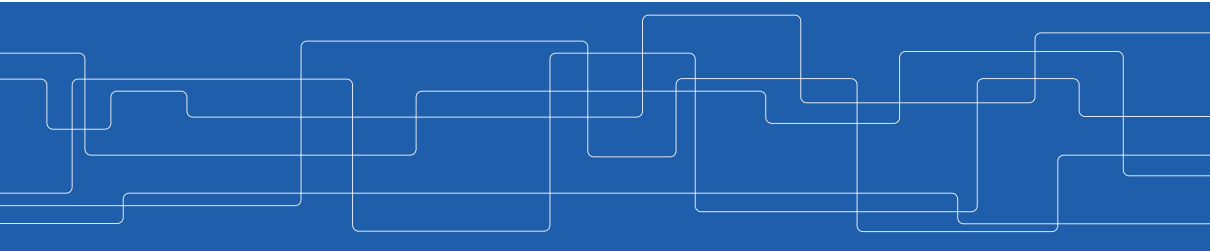
PaRISian particle Gibbs samplers for state and parameter learning in nonlinear state-space models

Jimmy Olsson
jimmyol@kth.se

In collaboration with G. Cardoso, S. Le Corff, Y. Janati El Idrissi, E. Moulines, S. Samsonov, and A. Thin



Papers of relevance

- G. Cardoso, S. Samsonov, A. Thin, E. Moulines, and J. Olsson. "BR-SNIS: Bias-reduced self-normalized importance sampling". In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- G. Cardoso, E. Moulines, and J. Olsson. "Particle-based, rapid incremental smoother meets particle Gibbs". To appear in *Statistica Sinica*, 2023.
- G. Cardoso, Y. Janati El Idrissi, S. Le Corff, E. Moulines, and J. Olsson. "State and parameter learning with PaRIS particle Gibbs". To be presented at the 2023 International Conference on Machine Learning (ICML).

This talk

Bias-reduced self-normalised importance sampling (`BR-SNIS`)

PaRISian particle Gibbs (`PPG`)

This talk

Bias-reduced self-normalised importance sampling (`BR-SNIS`)

PaRISian particle Gibbs (`PPG`)

Self-normalised importance sampling (SNIS)

- We aim to sample from $\pi(dx) \propto w(x)\,\lambda(dx)$ (the target) on some state space $(\mathbb{X}, \mathcal{X})$, where w is some positive weight function and λ some instrumental distribution (the proposal) on $(\mathbb{X}, \mathcal{X})$.
- The SNIS estimator [Gew89] of $\pi h = \int h(x)\,\pi(dx)$, h being a π-integrable objective function, is given by

$$\Pi_M h(\boldsymbol{\xi}) = \sum_{i=1}^{M} \frac{w(\xi^i)}{\sum_{j=1}^{M} w(\xi^j)} h(\xi^i),$$

where $\boldsymbol{\xi} = (\xi^1, \ldots, \xi^M) \sim \lambda^{\otimes M}$.

Bias and MSE of SNIS

- As shown in [Aga+17], for every M and h such that $\|h\|_\infty \le 1$,

$$|\mathbb{E}[\Pi_M h(\boldsymbol{\xi})] - \pi h| \le (12/M)\kappa[\pi, \lambda] \qquad \text{(bias)}$$

$$\mathbb{E}\left[\{\Pi_M h(\boldsymbol{\xi}) - \pi h\}^2\right] \le \underbrace{(4/M)\kappa[\pi, \lambda]}_{\mathsf{MSE}_M^{snis}} \qquad \text{(MSE)}$$

  where $\kappa[\pi, \lambda] \coloneqq \lambda(w^2)/\lambda^2(w)$.

- *In this talk*, we
  - show how a bias-reduced modification of SNIS can be obtained at the cost of a controllable increase of MSE by shuffling randomly the samples $\boldsymbol{\xi}$.
  - furnish the same with rigorous error bounds.

Prelude: iterated sampling importance resampling (i-SIR)

- The i-SIR algorithm [Tje04] (see also [ADH10]) generates a Markov chain $(\upsilon_t)_t$ on $\mathbb{X}$ as follows: given υ_t,

  draw ι uniformly over $\{1, \ldots, N\}$;
  set $\xi_{t+1}^{\iota} \leftarrow \upsilon_t$;
  for $i \in \{1, \ldots, N\} \setminus \iota$ **do**
  draw $\xi_{t+1}^{i} \sim \lambda$;
  end
  set $\kappa \leftarrow i$ with probability $\propto w(\xi_{t+1}^{i})$;
  set $\upsilon_{t+1} \leftarrow \xi_{t+1}^{\kappa}$;

- The chain $(\upsilon_t)_t$ can be shown to allow π as a stationary distribution.

i-SIR as a Gibbs sampler

- Note that i-SIR can be described by the two-stage procedure

$$(v_t, \boldsymbol{\xi}_t) \xrightarrow{\Lambda_N} (v_t, \boldsymbol{\xi}_{t+1}) \xrightarrow{\Pi_N} (v_{t+1}, \boldsymbol{\xi}_{t+1}),$$

  where $\Lambda_N(v_t, d\boldsymbol{x}) := \frac{1}{N} \sum_{i=1}^N \delta_{v_t}(dx^i) \prod_{j \neq i} \lambda(dx^j)$.

- Now, there exists a probability distribution $\boldsymbol{\lambda}_N$ on $\mathbb{X}^N$ such that

  (i) $\varphi_N(d(y, \boldsymbol{x})) := \pi(dy)\, \Lambda_N(y, d\boldsymbol{x}) = \boldsymbol{\lambda}_N(d\boldsymbol{x})\, \Pi_N(\boldsymbol{x}, dy)$ (dual representation)

  (ii) $\int \Pi_N h(\boldsymbol{x})\, \varphi_N(d\boldsymbol{x}) = \pi h.$ (unbiasedness)

- Thus, i-SIR can be embedded into a systematic-scan Gibbs sampler targeting the distribution φ_N, under which $\Pi_N h(\boldsymbol{\xi})$ is unbiased.

Bias-reduced SNIS (BR-SNIS)

- Thus, let us consider an estimator formed as an average across $(\Pi_N h(\xi_t))_t$.
- Thus, being ready to generate $M = (N-1)m$ samples ξ from λ, run i-SIR for m iterations, producing mini-batches $(\xi_t)_{t=1}^m$, and return

$$\Pi_{N,(m_0,m)} h(\xi) \coloneqq (m - m_0)^{-1} \sum_{t=m_0+1}^{m} \Pi_N h(\xi_t),$$

  where $m_0 < m$ is some burn-in.
- Wrapper requiring only random shuffling of ξ!
- Similar ideas appear in, *e.g.*, [Tje04; ADH10].

Convergence of BR-SNIS

- Recall that $\Pi_{N,(m_0,m)} h(\boldsymbol{\xi}) = (m - m_0)^{-1} \sum_{t=m_0+1}^{m} \Pi_N h(\boldsymbol{\xi}_t)$.
- letting $M = (N-1)m$ and $\phi := (m - m_0)/m$, yields $M\phi$ effective samples.

Theorem ([Car+22])

Assume that $\|w\|_\infty < \infty$. Then for all $N \geq 2$ there exist $\kappa_N \in (0,1)$, $\zeta^{bias} > 0$, and $\zeta^{hpd} > 0$ such that for every initial distribution $\boldsymbol{\mu}$, h such that $\|h\|_\infty \leq 1$, and $m_0 < m$,

(i) $\left|\mathbb{E}_{\boldsymbol{\mu}}\left[\Pi_{N,(m_0,m)} h(\boldsymbol{\xi})\right] - \pi h\right| \leq \zeta^{bias} \kappa_N^{m_0}/(M\phi)$

(ii) $\mathbb{E}_{\boldsymbol{\mu}}\left[\left\{\Pi_{N,(m_0,m)} h(\boldsymbol{\xi}) - \pi h\right\}^2\right] \leq \mathrm{MSE}^{snis}_{M\phi} + o\{1/(M\phi)\}$

(iii) *for every $\delta \in (0,1)$, $|\Pi_{N,(m_0,m)} h(\boldsymbol{\xi}) - \pi h| \leq \zeta^{hpd}\{\ln(4/\delta)\}^{1/2}/(M\phi)^{1/2}$ with probability at least $1 - \delta$.*

Variance reduction by sample permutation

- Increasing m_0 reduces bias but increases MSE; however, to control the latter we may proceed as follows:

  for $b = 1 \to B$ **do**
  > generate a random permutation $\boldsymbol{\xi}^{(b)}$ of $\boldsymbol{\xi}$;
  > run `i-SIR` on $\boldsymbol{\xi}^{(b)}$;
  > compute the `BR-SNIS` estimator $\Pi_{N,(m_0,m)} h(\boldsymbol{\xi}^{(b)})$;

  end

  return $\Pi^B_{N,(m_0,m)} h(\boldsymbol{\xi}) := B^{-1} \sum_{b=1}^B \Pi_{N,(m_0,m)} h(\boldsymbol{\xi}^{(b)})$;

- Here $\Pi^B_{N,(m_0,m)} h(\boldsymbol{\xi})$ and $\Pi_{N,(m_0,m)} h(\boldsymbol{\xi})$ have the same bias.
- Letting, *e.g.*, $m_0 = m - 1$ and $B = m$ entails an $O(1/M)$ variance.

Gaussian mixture model

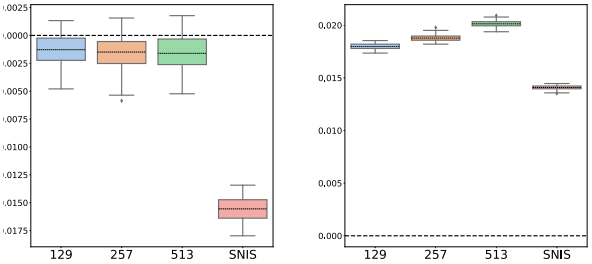


Figure: Comparison of bias (left panel) and MSE (right panel) of `BR-SNIS` and `SNIS` for the fixed budget $M = 16{,}384$, mini-batch sizes $N \in \{129, 257, 513\}$, $m_0 = m - 1$, and $B = m$. Here π is a mixture of two 7-dimensional Gaussian distributions, λ is a Student's t-distribution ($\nu = 3$), and $h = 1_A - 1_B$.

Application to Bayesian logistic regression

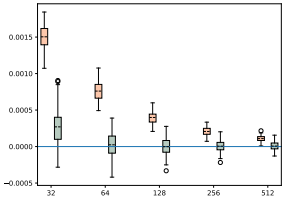
- Let $\mathcal{D} \coloneqq (\boldsymbol{x}_i, y_i)_{i=1}^n$ be covariates in $\mathbb{R}^d$ and binary responses in $\{-1, 1\}$.
- Let $p_\theta(y_i \mid \boldsymbol{x}_i) \coloneqq 1/\{1 + \exp(-\boldsymbol{x}_i^\top \theta y_i)\}$ be the likelihood of y_i given $\boldsymbol{x}_i$ and π_0 a Gaussian prior on $\Theta \subseteq \mathbb{R}^d$.
- In this case, the posterior is

$$\pi(d\theta \mid \mathcal{D}) \propto \exp(\ell_n(\theta))\, \pi_0(d\theta),$$

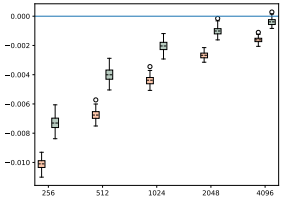
  where $\ell_n(\theta) = \sum_{i=1}^n \ln p_\theta(y_i \mid \boldsymbol{x}_i)$.
- In this setting, we use `BR-SNIS` to estimate

$$\mathbb{E}[\theta_j \mid \mathcal{D}] = \int \theta_j \, \pi(d\theta_j \mid \mathcal{D}).$$

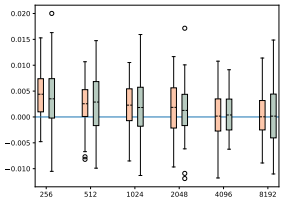
Application to Bayesian logistic regression (cont'd)



(a) HFCR, θ_8



(b) WDBC, θ_{11}



(c) CT, θ_6

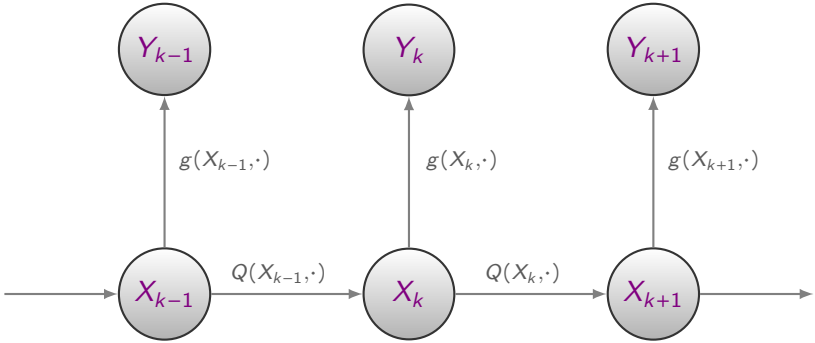
Figure: Estimated biases of approximations Bayes's estimator of θ_j obtained with BR-SNIS (■) and SNIS (■) for different budgets M and different data sets HFCR = *Heart Failure Clinical Records* ($d = 13$, $n = 299$), WDBC = *Wisconsin Diagnostic Breast Cancer* ($d = 30$, $n = 569$), and CT = *Cover Type* ($d = 55$, $n = 40{,}000$).

This talk

Bias-reduced self-normalised importance sampling (BR-SNIS)

PaRISian particle Gibbs (PPG)

General state-space models (SSM)



- An SSM is a partially observed Markov chain $(X_k, Y_k)_k$ on $\mathbb{X} \times \mathbb{Y}$ characterised by (i) a state transition kernel Q (with density q) , (ii) an emission density g, and (iii) an initial distribution μ.

Additive smoothing in SSM

- Our aim is to approximate $\pi_{0:n}h = \mathbb{E}_\mu[h(X_{0:n}) \mid Y_{0:n-1}]$, where
  - $\pi_{0:n}(dx_{0:n}) \propto \mu(dx_0) \prod_{k=0}^{n-1} g(x_k, y_k)\, Q(x_k, dx_{k+1})$,
  - $h(x_{0:n}) = \sum_{k=0}^{n-1} h_k(x_k, x_{k+1})$.
- This can be done by an analogous two-stage procedure

$$(\upsilon_t, \boldsymbol{\xi}_t) \xrightarrow{\Lambda_N} (\upsilon_t, \boldsymbol{\xi}_{t+1}) \xrightarrow{\Pi_N} (\upsilon_{t+1}, \boldsymbol{\xi}_{t+1}),$$

  where, in this case,
  - $\upsilon_t = (\upsilon_0, \ldots, \upsilon_n)_t$ is a random *path* in $\mathbb{X}^{n+1}$
  - $\boldsymbol{\xi}_t$ contains N paths and associated real-valued statistics $((\xi_{0|n}, \ldots, \xi_{n|n})^i, \beta_n^i)_t$ such that $\bar{\boldsymbol{\beta}}_t \coloneqq \frac{1}{N}\sum_{i=1}^N \beta_{n,t}^i$ approximates $\pi_{0:n}h$.

PaRISian particle Gibbs (PPG)

- The operation $\boldsymbol{\xi} \sim \Lambda_N(\upsilon, \cdot)$ combines *conditional SMC* [ADH10] and the *particle-based, rapid incremental smoother* (PaRIS) [OW17] by
  - evolving a particle cloud $(\xi_k^i)_{i=1}^N$ conditionally on $\upsilon = (\upsilon_0, \ldots, \upsilon_n)$.
  - letting recursively, for $k \in \{0, \ldots, n-1\}$,

$$\beta_{k+1}^i \leftarrow \frac{1}{J} \sum_{j=1}^{J} \left(\beta_k^{I^j} + h_k(\xi_k^{I^j}, \xi_{k+1}^i) \right),$$

$$(\xi_{0|k+1}, \ldots, \xi_{k+1|k+1})^i \leftarrow ((\xi_{0|k}, \ldots, \xi_{k|k})^{I^1}, \xi_{k+1}^i),$$

  where $I^j \sim cat((g(\xi_k^\ell, y_k) q(\xi_k^\ell, \xi_{k+1}^i))_{\ell=1}^N)$ and $2 \le J \ll N$.

- The operation $\upsilon \sim \Pi_N(\boldsymbol{\xi}, \cdot)$ resamples uniformly among the paths $(\xi_{0|n}, \ldots, \xi_{n|n})^i$ in $\boldsymbol{\xi}$.

Convergence of PPG

- As before, define $\Pi_{(m_0,m),N} h(\boldsymbol{\xi}) \coloneqq (m - m_0)^{-1} \sum_{t=m_0+1}^{m} \bar{\boldsymbol{\beta}}_t$.

Theorem ([CMO23])

Under certain strong mixing assumptions there exist, for all $N \geq 2$ and $J \geq 2$, $\kappa_N \in (0,1)$, $\zeta_n^{bias} > 0$, and $\zeta_n^{mse} > 0$ such that for every $\boldsymbol{\mu}$, h, and $m_0 < m$,

(i) $\left| \mathbb{E}_{\boldsymbol{\mu}} \left[\Pi_{(m_0,m),N} h(\boldsymbol{\xi}) \right] - \pi_{0:n} h \right| \leq \zeta_n^{bias} \kappa_N^{m_0} \left(\sum_{k=0}^{n-1} \| \tilde{h}_k \|_\infty \right) / \{(1 - \kappa_N) M \phi\}$

(ii) $\mathbb{E}_{\boldsymbol{\mu}} \left[\{ \Pi_{(m_0,m),N} h(\boldsymbol{\xi}) - \pi_{0:n} h \}^2 \right] \leq \zeta_n^{mse} \left(\sum_{k=0}^{n-1} \| \tilde{h}_k \|_\infty \right)^2 / \{(1 - \kappa_N) M \phi\} + o\{1/(M\phi)\}.$

Application to standard SSM

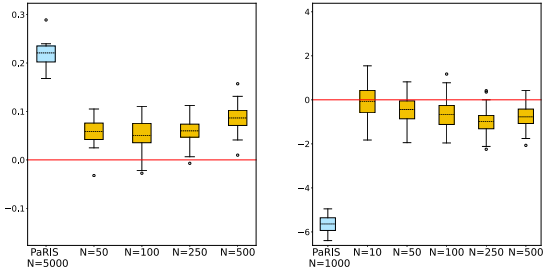


Figure: PaRIS and PPG bias dispersions for a linear Gaussian (left panel) and a stochastic volatility (right panel) model as a function of the mini-batch size N for fixed computational budgets $M = Nm$ of 5,000 (linear Gaussian) and 1,000 (stochastic volatility) and with $m_0 = \lfloor 2^{-1} m \rfloor$ burn-in steps.

PPG-based score ascent

- We wish to estimate some model parameter $\theta \in \Theta$ given data $y_{0:n}$.
- In order to find a zero of the score $s_n(\theta) = \pi_{0:n,\theta} h_\theta$, where $h_\theta(x_{0:n})$ is the gradient of $\theta \mapsto \sum_{k=0}^{n-1} \log\{g_\theta(x_k, y_k) q_\theta(x_k, x_{k+1})\}$, we run L iterations of

$$\theta_{\ell+1} = \theta_\ell + \gamma_{\ell+1} \Pi_{(m_0,m),N} h_{\theta_\ell}(\boldsymbol{\xi}).$$

Theorem ([Car+23])

Under certain assumptions there exist $a_{N,n} > 0$ and $b_{N,n} > 0$ such that for all L,

$$\mathbb{E}[\|s_n(\theta_\lambda)\|^2] \leq \frac{c_{N,n} + d_{N,n} \sum_{\ell=0}^{L} \gamma_\ell^2}{\sum_{\ell=1}^{L} \gamma_\ell},$$

with $\lambda \sim cat((\gamma_\ell)_{\ell=0}^{L})$. Letting $\gamma_\ell \propto 1/\sqrt{\ell}$ yields $\mathbb{E}[\|s_n(\theta_\lambda)\|^2] = O(\log L/\sqrt{L})$.

PPG-based score ascent (cont'd)

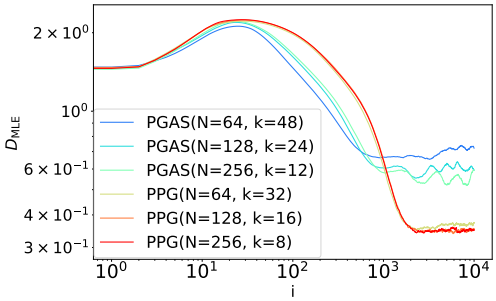


Figure: Plot of L_2 distances to the MLE estimator of the keofficients of a linear Gaussian SSM as a function of L for different configurations of the PPG. Here $n = 999$. Solid lines and shaded regions correspond to means and approximate confidence intervals obtained with 25 replicates.

Conclusions

- Calculating $\Pi^{B}_{N,(m_0,m)} h(\boldsymbol{\xi})$ requires only computationally cheap random shuffling of $(\xi^i, w(\xi^i), h(\xi^i))_{i=1}^{M}$ (the generation of which constitutes the computational bottleneck of `SNIS`).
- `BR-SNIS` can be applied off-the-shelf whenever `SNIS` is to be used.
- A similar bias-reduced estimator, `PPG`, can be obtained for SSM using the `PaRIS` estimator.
- The bias reduction provided by the `PPG` is obtained at the cost of an increase of MSE that cannot be easily controlled.

References I

C. Andrieu, A. Doucet, and R. Holenstein. "Particle Markov chain Monte Carlo methods". In: *J. R. Stat. Soc. Ser. B* 72.3 (2010), pp. 269–342.

S. Agapiou et al. "Importance Sampling: Intrinsic Dimension and Computational Cost". In: *Statistical Science* 32.3 (2017), pp. 405–431.

G. Cardoso et al. "BR-SNIS: bias-reduced self-normalized importance sampling". In: *Advances in Neural Information Processing Systems (NeurIPS '22)*. 2022.

G. Cardoso et al. *State and parameter learning with PaRIS particle Gibbs*. Tech. rep. 2023.

G. Cardoso, E. Moulines, and J. Olsson. "Particle-based, rapid incremental smoother meets particle Gibbs". In: *Statistica Sinica* (2023). To appear.

References II

J. Geweke. "Bayesian inference in econometric models using Monte Carlo integration". In: *Econometrica* 57.6 (1989), pp. 1317–1339.

J. Olsson and J. Westerborn. "Efficient particle-based online smoothing in general hidden Markov models: The PaRIS algorithm". In: *Bernoulli* 23.3 (2017), pp. 1951–1996.

H. Tjelmeland. *Using all Metropolis–Hastings proposals to estimate mean values*. Tech. rep. 2004.