Sequential Core-Set Monte Carlo (Supplementary Material)

Boyan Beronov[1] **Christian Weilbach**[1] **Frank Wood**[1] **Trevor Campbell**[2]

[1]Dept. of Computer Science, University of British Columbia, Vancouver, Canada
[2]Dept. of Statistics, University of British Columbia, Vancouver, Canada

1 ALGORITHMS

1.1 SEQUENTIAL MONTE CARLO

Algorithm 1 Resample-move SMC [Chopin, 2002, Gilks and Berzuini, 2001] for data tempering in exchangeable models

1: **procedure** SMC(*population size* K, *proc.* REJUV)
▷ *initialize particles and memory*
2: $(\pi_k^{(0)}, \theta_k^{(0)}) \leftarrow$ Equation (2), $\quad M^{(0)} \leftarrow 0$
3: **for** $t \leftarrow 1, \dots, T$ **do**
▷ *update particle weights with new data batch* $(x_b^{(t)})_{b=1}^B$
4: $(\bar{\pi}_k^{(t)})_{k=1}^K \leftarrow$ Equation (3)
▷ *expand memory with new data batch* $(x_b^{(t)})_{b=1}^B$
5: $M^{(t)} \leftarrow M^{(t-1)} + B, \quad (1, \bar{u}_j^{(t)})_{j=1}^{M^{(t)}} \leftarrow$ Equation (11)
▷ *resample particles*
6: **if** ESS$(\bar{\pi}^{(t)}) <$ threshold **then**
7: $(\widehat{\pi}_k^{(t)}, \widehat{\theta}_k^{(t)})_{k=1}^K \leftarrow$ Equation (5)
8: **else**
9: $(\widehat{\pi}_k^{(t)}, \widehat{\theta}_k^{(t)})_{k=1}^K \leftarrow (\bar{\pi}_k^{(t)}, \theta_k^{(t-1)})_{k=1}^K$
10: **end if**
▷ *rejuvenate particles using memory (Algorithm 2)*
11: $(\pi_k^{(t)}, \theta_k^{(t)})_{k=1}^K \leftarrow \texttt{REJUV}\Big((\widehat{\pi}_k^{(t)}, \widehat{\theta}_k^{(t)})_{k=1}^K, (1, \bar{u}_j^{(t)})_{j=1}^{M^{(t)}}\Big)$
12: **end for**
13: **return** *particles* $(\pi_k^{(T)}, \theta_k^{(T)})_{k=1}^K$
14: **end procedure**

For completeness, we provide in Algorithm 1 the original resample-move SMC method for sequential inference in static models, using the notation in this work. In comparison to SCMC (Algorithm 1), it has uniform data weights in line 5, lacks the core-set projection in lines 12-17, and does not produce a core-set memory as an output.

1.2 WEIGHTED METROPOLIS-HASTINGS FOR CORE-SET REJUVENATION

The simplest possible rejuvenation method, used inside Algorithm 1 for all experiments, is the Metropolis-Hastings kernel in Algorithm 2. The only difference to the standard procedure is the non-uniform weighting of data likelihoods in line 4, cf. Section 3.2. See Section 2 for details on the choice of proposal kernels q in individual experiments.

Algorithm 2 Weighted rejuvenation kernel using Metropolis-Hastings

procedure REJUV(*population* $(\widehat{\pi}_k^{(t)}, \widehat{\theta}_k^{(t)})_{k=1}^K$, *memory* $(\bar{w}_j^{(t)}, \bar{u}_j^{(t)})_{j=1}^{\bar{C}^{(t)}}$, *steps* J, *proposal* q)
  $\forall k \in [K] : \theta_k^{(0)} \leftarrow \widehat{\theta}_k^{(t)}$
  for $j \leftarrow 1, \ldots, J$ **do**
    ▷ *propose, compute acceptance probability (Equation* (18)*)*
    $\forall k \in [K] : \theta'_k \sim q(\cdot \,|\, \theta_k^{(j-1)}), \quad a_k^{(j)} \leftarrow \alpha(\theta'_k, \theta_k^{(j-1)})$
    ▷ *accept / reject*
    $\forall k \in [K] : \theta_k^{(j)} \sim a_k^{(j)} \cdot \delta_{\theta'_k} + (1 - a_k^{(j)}) \cdot \delta_{\theta_k^{(j-1)}}$
  end for
  return $(\widehat{\pi}_k^{(t)}, \theta_k^{(J)})_{k=1}^K$
end procedure

1.3 GREEDY ITERATIVE GEODESIC ASCENT

Since Hilbert core-set projection can be reduced to the standard sparse nonnegative least-squares (SNNLS) problem in Equation (16), SCMC (Algorithm 1) is parametrized by the choice of an SNNLS solver. In this work, we choose GIGA [Campbell and Broderick, 2018] for this purpose, because it was designed specifically for this application, and because it provides the currently best accuracy-cost trade-off. Bayesian core-set methods such as [Campbell and Beronov, 2019] can reach higher levels of precision, but require their own internal sequential inference procedures with significantly higher runtimes.

We note that prior work on model-agnostic Bayesian core-set construction was limited to the batch setting [Huggins et al., 2016, Campbell and Broderick, 2018, 2019, Campbell and Beronov, 2019], in which the input data have uniform weights and the target vector in Equation (15) is given by $b = A \cdot \vec{1}$. The iterative core-set construction task performed by CPF instead requires the solution of an SNNLS problem with general weights as in Equation (16), and we provide the natural generalization of GIGA to this scenario in Algorithm 3.

2 EXPERIMENT DETAILS

In this section, we provide analytical background on the models and the parameters for each experiment. The hyper-parameters for rejuvenation kernels are listed in Table 1, and all experiments are conducted with 3 rejuvenation steps. Following the concentration rate of a conjugate Gaussian posterior, we reduce the variance of proposal kernels linearly in the number of data points, where we denote the total number of observations at SMC step t by σ_t.

2.1 AUTO-REGRESSIVE PROCESS

2.1.1 Rejuvenation Proposal

We use the same Gaussian kernel as in Equation (6), with parameters listed in Table 1 and $\Sigma = 3$. We have done no tuning and, given the simplicity of the experiment, assume that many other parameters would work as well.

2.2 NORMAL-INVERSE-WISHART

When the generative model constitutes an exponential family, the space of core-set posteriors generally lies within the same family. Our first experiment makes use of this circumstance in order to directly evaluate the approximation error of CPF in Figure 3, without introducing an additional inference problem.

Algorithm 3 Greedy iterative geodesic ascent—generalized from $b = A \cdot \vec{1}$ in [Campbell and Broderick, 2018]

procedure GIGA(*dictionary* $A \in \mathbb{R}^{K \times N}$, *target* $b \in \mathbb{R}^K$, *core-set size* M, *tolerance* ε)
  ▷ *normalize vectors and initialize weights to 0*
  $\beta \leftarrow \frac{b}{\|b\|_2}, \quad \forall\, n \in [N]: \quad \alpha_n \leftarrow \frac{A \cdot e_n}{\|A \cdot e_n\|_2}$
  $w_0 \leftarrow 0, \quad \beta_0 \leftarrow 0, \quad \epsilon_0 \leftarrow \|b\|_2$
  for $t \in \{0, \ldots, M-1\}$ **do**
    ▷ *compute the geodesic direction for each data point*
    $d_t \leftarrow \frac{\beta - \langle \beta, \beta_t \rangle \beta_t}{\left\|\beta - \langle \beta, \beta_t \rangle \beta_t\right\|_2}, \quad \forall n \in [N]: \; d_{t,n} \leftarrow \frac{\alpha_n - \langle \alpha_n, \beta_t \rangle \beta_t}{\left\|a_n - \langle \alpha_n, \beta_t \rangle \beta_t\right\|_2}$
    ▷ *choose the best geodesic*
    $n_t \leftarrow \operatorname{argmax}_{n \in [N]} \langle d_t, d_{t,n} \rangle$
    $\zeta_0 \leftarrow \langle \beta, a_{n_t} \rangle, \quad \zeta_1 \leftarrow \langle \beta, \beta_t \rangle, \quad \zeta_2 \leftarrow \langle a_{n_t}, \beta_t \rangle$
    ▷ *compute the step size*
    $\gamma_t \leftarrow \frac{\zeta_0 - \zeta_1 \zeta_2}{(\zeta_0 - \zeta_1 \zeta_2) + (\zeta_1 - \zeta_0 \zeta_2)}$
    ▷ *update the core-set*
    $w_{t+1} \leftarrow \frac{(1-\gamma_t) w_t + \gamma_t 1_{n_t}}{\left\|(1-\gamma_t)\beta_t + \gamma_t \alpha_{n_t}\right\|_2}, \quad \beta_{t+1} \leftarrow \frac{(1-\gamma_t)\beta_t + \gamma_t a_{n_t}}{\left\|(1-\gamma_t)\beta_t + \gamma_t \alpha_{n_t}\right\|_2}$
    $\epsilon_{t+1} \leftarrow \|A \cdot w_{t+1} - b\|_2$
    ▷ *check improvement in numerical precision*
    if $\epsilon_{t+1} > \epsilon_t \cdot (1 + \varepsilon)$ **then**
      $w_M \leftarrow w_t, \quad \beta_M \leftarrow \beta_t$
      break
    end if
  end for
  ▷ *scale the weights optimally*
  $\forall\, n \in [N]: \quad (w_M)_n \leftarrow (w_M)_n \cdot \frac{\|b\|_2}{\|A \cdot e_n\|_2} \langle \beta, \beta_M \rangle$
  return w_M
end procedure

2.2.1 Generative Model

The prior and likelihood of the NIW model read

$$p_0(m,\Sigma) = p\left(m,\Sigma \mid \mu_0,\sigma_0,\Psi_0,\nu_0\right) = \mathcal{N}\left(m \mid \mu_0, \frac{\Sigma}{\sigma_0}\right)\, \mathcal{W}^{-1}(\Sigma \mid \Psi_0,\nu_0) \tag{1}$$

$$p\left(x \mid m,\Sigma\right) = \mathcal{N}(x \mid m,\Sigma)\ . \tag{2}$$

2.2.2 Core-Set Posterior

The conjugacy of the standard NIW model generalizes to the case of weighted observations: Denoting by $|w| \coloneqq \sum_n w_n$ and $\overline{x} \coloneqq \frac{1}{|w|}\sum_n w_n \cdot x_n$ the total weight and the weighted mean of observations, the sufficient statistics of the exact posterior p_1 after observing the weighted data batch $(w_n, x_n)_n$ are

$$\sigma_1 = \sigma_0 + |w| \qquad \nu_1 = \nu_0 + |w| \qquad \mu_1 = \frac{\sigma_0 \cdot \mu_0 + |w| \cdot \overline{x}}{\sigma_0 + |w|} \tag{3}$$

$$\Psi_1 = \Psi_0 + \frac{\sigma_0 \cdot |w|}{\sigma_0 + |w|}\left(\mu_0 - \overline{x}\right)\left(\mu_0 - \overline{x}\right)^T + \sum_n w_n \left(x_n - \overline{x}\right)\left(x_n - \overline{x}\right)^T\ . \tag{4}$$

This can be shown as follows: Using $\langle x,y\rangle_\Sigma \coloneqq x^T\Sigma^{-1}y$ and $\|x\|_\Sigma^2 \coloneqq \langle x,x\rangle_\Sigma$ for the inner product and norm induced by Σ,

$$\begin{aligned}
&\left(\frac{p_0(\cdot)}{p_1(\cdot)} \cdot \prod_n p(x_n \mid \cdot)^{w_n}\right)(m,\Sigma) \\
&\overset{(1),(2)}{=} \frac{\mathcal{N}\left(m \,|\, \mu_0, \frac{\Sigma}{\sigma_0}\right)}{\mathcal{N}\left(m \,|\, \mu_1, \frac{\Sigma}{\sigma_1}\right)} \cdot \frac{\mathcal{W}^{-1}\left(\Sigma \,|\, \Psi_0,\nu_0\right)}{\mathcal{W}^{-1}\left(\Sigma \,|\, \Psi_1,\nu_1\right)} \cdot \exp\left\{\sum_n w_n \cdot \log \mathcal{N}(x_n \,|\, m,\Sigma)\right\} \\
&\overset{\mathcal{N},\mathcal{W}}{\propto} \frac{\exp\left\{-\frac{1}{2}\left[\ln|\frac{\Sigma}{\sigma_0}| + \|m-\mu_0\|^2_{\Sigma/\sigma_0}\right]\right\}}{\exp\left\{-\frac{1}{2}\left[\ln|\frac{\Sigma}{\sigma_1}| + \|m-\mu_1\|^2_{\Sigma/\sigma_1}\right]\right\}} \cdot \frac{\exp\left\{-\frac{1}{2}\left[(\nu_0+D+1)\ln|\Sigma| + \mathrm{Tr}\big(\Psi_0\Sigma^{-1}\big)\right]\right\}}{\exp\left\{-\frac{1}{2}\left[(\nu_1+D+1)\ln|\Sigma| + \mathrm{Tr}(\Psi_1\Sigma^{-1})\right]\right\}} \\
&\qquad \cdot \exp\left\{-\frac{1}{2}\left[|w|\cdot\ln|\Sigma| + \sum_n w_n \|x_n - m\|^2_\Sigma\right]\right\} \\
&\overset{(3)}{\propto} \exp\left\{\mathrm{Tr}\big(\left[\Psi_1 - \Psi_0\right]\Sigma^{-1}\big) - \sum_n w_n \|x_n - m\|^2_\Sigma - \sigma_0\|m-\mu_0\|^2_\Sigma + \sigma_1\|m-\mu_1\|^2_\Sigma\right\}^{\frac{1}{2}} \\
&\overset{(3)}{=} \exp\left\{\mathrm{Tr}\big(\left[\Psi_1 - \Psi_0\right]\Sigma^{-1}\big) - |w|\,\|m\|^2_\Sigma + 2\,\langle m,\, |w|\,\overline{x}\rangle_\Sigma - \sum_n w_n\|x_n\|^2_\Sigma - \sigma_0\|m-\mu_0\|^2_\Sigma\right. \\
&\qquad\qquad \left. + (\sigma_0+|w|)\left\|m - \frac{\sigma_0\cdot\mu_0 + |w|\cdot\overline{x}}{\sigma_0+|w|}\right\|^2_\Sigma\right\}^{\frac{1}{2}} \\
&\overset{(5)}{=} \exp\left\{\sigma_0\left(\mathrm{Tr}\big(\mu_0\mu_0^T\Sigma^{-1}\big) - \|\mu_0\|^2_\Sigma\right) - \sigma_1\left(\mathrm{Tr}\big(\mu_1\mu_1^T\Sigma^{-1}\big) - \|\mu_1\|^2_\Sigma\right) + \sum_n w_n\left(\mathrm{Tr}\big(x_n x_n^T\Sigma^{-1}\big) - \|x_n\|^2_\Sigma\right)\right\}^{\frac{1}{2}} \\
&\overset{\mathrm{Tr}}{=} 1\,,
\end{aligned}$$

Experiment	α_σ	α_ν	β
NIW	2	2	1.015
BLR	0.002	–	1
AR(1)	1.5	–	1

Table 1: Parameters for the rejuvenation kernels.

where the penultimate step uses the following identity:

$$\begin{aligned}
\Psi_1 - \Psi_0 &\overset{(4)}{=} \frac{\sigma_0 \cdot |w|}{\sigma_0 + |w|} (\mu_0 - \overline{x}) (\mu_0 - \overline{x})^T - |w| \cdot \overline{x} \cdot \overline{x}^T + \sum_n w_n \left\{ x_n x_n^T + 2 \cdot \overline{x} \cdot \overline{x}^T - \left(x_n \overline{x}^T + \overline{x} x_n^T \right) \right\} \\
&= \frac{\sigma_0 \cdot |w|}{\sigma_0 + |w|} \left\{ \mu_0 \mu_0^T + \left(1 - \left(\frac{|w|}{\sigma_0} + 1 \right) \right) \overline{x} \cdot \overline{x}^T - \left(\mu_0 \overline{x}^T + \overline{x} \mu_0^T \right) \right\} + \sum_n w_n \cdot x_n x_n^T \\
&= \sigma_0 \cdot \mu_0 \mu_0^T - \frac{1}{\sigma_0 + |w|} \left\{ \sigma_0^2 \cdot \mu_0 \mu_0^T + |w|^2 \cdot \overline{x} \cdot \overline{x}^T + \sigma_0 |w| \left(\mu_0 \overline{x}^T + \overline{x} \mu_0^T \right) \right\} + \sum_n w_n \cdot x_n x_n^T \\
&\overset{(3)}{=} \sigma_0 \cdot \mu_0 \mu_0^T - \sigma_1 \cdot \mu_1 \mu_1^T + \sum_n w_n \cdot x_n x_n^T
\end{aligned} \tag{5}$$

□

2.2.3 Convergence Metric

Posterior convergence in this experiment was measured using the maximum mean discrepancy,

$$\mathrm{MMD}_k^2(\pi, \hat{\pi}) := \mathbb{E}_{x,x' \sim \pi}[\, k(x, x') \,] + \mathbb{E}_{y,y' \sim \hat{\pi}}[\, k(y, y') \,] - 2\, \mathbb{E}_{x \sim \pi, y \sim \hat{\pi}}[\, k(x, y) \,],$$

a reproducing kernel Hilbert space metric which is well-suited for comparing distributions in different representations [Gretton et al., 2012]. The kernel $k : \Theta \times \Theta \to [0, 1]$ in the space of Gaussian sufficient statistics was chosen as a Gaussian RBF, with a radius provided by the analytically computable Jeffrey divergence between model parameters,

$$k\Big(m, \Sigma \,|\, \hat{m}, \hat{\Sigma}\Big) \propto \exp\left\{ -\frac{1}{\alpha}\, \mathcal{D}_J \left(\mathcal{N}(m, \Sigma) \,\|\, \mathcal{N}(\hat{m}, \hat{\Sigma}) \right)^2 \right\},$$

and with a scale adjusted as $\alpha = \frac{1}{2}\, \mathcal{D}_J \left(p_0 \,\|\, p_T \right)$ in terms of the prior p_0 and the exact posterior p_T.

2.2.4 Rejuvenation Proposal

Given σ_t, ν_t according to (3) for the SMC step t, the Metropolis-Hastings proposal kernel used for rejuvenation in the NIW experiments is constructed as follows:

$$q_t \left(\hat{m}, \hat{\Sigma} \mid m, \Sigma ;\ \alpha_\sigma, \alpha_\nu, \beta \right) = \mathcal{N} \left(\hat{m} \mid m, \frac{\hat{\Sigma}}{\alpha_\sigma \sigma_t^\beta} \right) \mathcal{W}^{-1} \left(\hat{\Sigma} \mid (\alpha_\nu \nu_t^\beta - n - 1)\Sigma,\ \alpha_\nu \nu_t^\beta \right) .$$

2.3 LOGISTIC REGRESSION

Unless stated otherwise in Section 5.3, we use the same setup for our Bayesian logistic regression experiment as in [Huggins et al., 2016]. The rejuvenation proposal is a Gaussian kernel centred at each particle,

$$q_t \left(\cdot \mid \Sigma ;\ \alpha_\sigma, \beta \right) = \mathcal{N} \left(\cdot \mid 0, \frac{\Sigma}{\alpha_\sigma \sigma_t^\beta} \right) , \tag{6}$$

where the tuning parameters are listed in Table 1, and clipping is introduced to ensure that $\alpha_\sigma \sigma_t^\beta > 1.0$ for all t. To calibrate all SMC variants equally well, Σ is chosen by estimation from $10,000$ posterior samples of a STAN [Carpenter et al., 2017] oracle. We set only two control parameters for STAN explicitly: `adapt_delta` $= 0.9$ and `max_treedepth` $= 15$.