Sampling from "intractable" posterior distributions: some basic algorithms

Anirban Bhattacharya

http://www.stat.tamu.edu/~anirbanb

March 8, 2018

1 Motivation

Denote the observations by $\mathbf{x}$ with sampling density (pdf/pmf) $f(\mathbf{x} \mid \theta)$ with $\theta \in \Theta \subset \mathbb{R}^d$. Let $\pi(\cdot)$ be a prior on θ. The posterior

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{m(\mathbf{x})}, \quad m(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\pi(\theta)d\theta.$$

In the i.i.d. case, $f(\mathbf{x} \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$.

In non-conjugate settings, we need a general set of tools to "compute" the posterior density. Here, by "computing" the posterior density, we mean that we should be able to calculate any posterior functional, such as the posterior mean, variance, median, quantiles of θ or of $\psi(\theta)$, where ψ is a known function. Even in conjugate settings, we have seen examples where the posterior distribution of certain $\psi(\theta)$s may be hard to obtain analytically, and we had to resort to Monte Carlo techniques. In general, our aim is to be able to (approximately) sample from the posterior distribution, so that the distribution of any posterior functional can be approximated. For example, if $\theta_1, \ldots, \theta_T$ are (approximately) independent samples from the posterior, then $\psi(\theta_1), \ldots, \psi(\theta_T)$ are samples from the posterior distribution of $\psi(\theta) \mid \mathbf{x}$, and we can use these samples to approximate the posterior mean/median/quantiles etc of $\psi(\theta)$.

The main bottleneck in sampling from the posterior is that the normalizing constant $m(\mathbf{x})$ is generally "intractable". This may be due to the fact that the integral is not analytically available, or the integral is highly expensive to compute, or a combination of both. For example, if $f(x \mid \theta) \propto [1 + (x - \theta)^2]^{-1}$, a Cauchy distribution with location θ, and $\theta \sim N(0, 1)$, then the integral is clearly not a standard one. As a second example, consider $x \mid \theta \sim 0.5N(\mu_1, 1) + 0.5N(\mu_2, 1)$, with $\mu_1, \mu_2 \sim N(0, 1)$ independently. Then,

$$f(\mathbf{x} \mid \theta) = 2^{-n} \sum_{j=0}^{n} \sum_{S:|S|=j} \Big[\int \prod_{i \in S} \phi(x_i - \mu_1)\phi(\mu_1)d\mu_1 \Big] \Big[\int \prod_{l \in S^c} \phi(x_l - \mu_2)\phi(\mu_2)d\mu_2 \Big],$$

where ϕ is the standard normal cdf and S denotes a subset of $\{1, \ldots, n\}$ with $|S|$ its size. Clearly, each of the inner integrals can be calculated analytically, but we have an outer sum over 2^n terms.

2 The Bootstrap filter

The Bootstrap filter is operationally very similar to self-normalized importance sampling we studied earlier. The additional observation here is that we can not only approximate integrals under the posterior, but also obtain a discrete approximation to the posterior, which is useful

for example in obtaining credible intervals.

Algorithm (Bootstrap filter):

(i) Sample $\theta_1, \ldots, \theta_T \sim \pi$ independently.

(ii) Set

$$w_t = \frac{f(\mathbf{x}_{obs} \mid \theta_t)}{\sum_{j=1}^{T} f(\mathbf{x}_{obs} \mid \theta_j)}.$$

(iii) Attach probability w_t to θ_t. In other words, $\hat{\Pi}_T := \sum_{t=1}^{T} w_t \delta_{\theta_t}$ is our (random) discrete approximation to the posterior distribution. Here and elsewhere, δ_u denotes a point mass at u.

It is straightforward to show that for any Borel set B, $\hat{\Pi}_T(B) \to \Pi(B \mid \mathbf{x}_{obs})$ as $T \to \infty$. To see this,

$$\hat{\Pi}_T(B) = \frac{T^{-1} \sum_{t=1}^{T} f(\mathbf{x}_{obs} \mid \theta_t) \mathbb{1}(\theta_t \in B)}{T^{-1} \sum_{t=1}^{T} f(\mathbf{x}_{obs} \mid \theta_t)} \to \frac{\int_B f(\mathbf{x}_{obs} \mid \theta) \pi(\theta) d\theta}{\int f(\mathbf{x}_{obs} \mid \theta) \pi(\theta) d\theta},$$

almost surely by SLLN. Clearly, the last expression is the posterior probability of the set B. Along similar lines, (and maybe with a few additional assumptions), we can show that for any "nice" function $g : \Theta \to \mathbb{R}$,

$$\sum_{j=1}^{T} w_j g(\theta_j) \to \int g(\theta) \pi(\theta \mid \mathbf{x}_{obs})$$

almost surely as $T \to \infty$, provided the right hand side exists and is finite. This in particular means we can approximate any posterior moments from the discrete approximation. Same is true of posterior quantiles, which allows us to construct credible intervals for the unknown parameters.

A useful modification of the Bootstrap filter can be achieved by sampling from an *importance density* q in the first step instead of the prior π. The weights then need to be appropriately adjusted to keep the target distribution the same. This importance density may be derived from a gaussian approximation to the posterior or a kernel density estimator fitted to a previous discrete approximation to the posterior.

Algorithm (Bootstrap filter with IS): Let q be a positive density on Θ.

(i) Sample $\theta_1, \ldots, \theta_T \sim q$ independently.

(ii) Set

$$\omega_t = \frac{f(\mathbf{x}_{obs} \mid \theta_t) \pi(\theta_t) / q(\theta_t)}{\sum_{j=1}^{T} f(\mathbf{x}_{obs} \mid \theta_j) \pi(\theta_j) / q(\theta_j)}.$$

(iii) Attach probability ω_t to θ_t, i.e., $\hat{\Pi}_T^{IS} := \sum_{t=1}^{T} \omega_t \delta_{\theta_t}$ is the discrete approximation to the posterior distribution.

Verify that all the properties of $\hat{\Pi}_T$ remain intact for $\hat{\Pi}_T^{IS}$. Indeed, with a "good" importance density q, $\hat{\Pi}_T^{IS}$ may be efficient by orders of magnitude. For choosing q, one thing to be careful about is that q is not too light tailed. If q has lighter tails than the posterior, then one may potentially underestimate uncertainty. A default choice is to use heavy tailed distributions like the t. The mean and covariance may be set to be the mle and a constant (> 1) multiple of the inverse Fisher information respectively in regular models.

The Bootstrap filter is ideal for low-dimensional problems; however, with increasing dimensions,

one needs more and more particles (i.e., larger T) to reliably estimate posterior quantities. One way to think about this is that there is a lot more "empty space" as dimension increases, so it is more likely that a random prior draw has very little posterior density. In such cases, one encounters something called "particle degeneracy", i.e., most of the particles have very small weight and all the weight concentrates on a very small number of particles. (If you are interested in a more mathematical description of this, there is a recent paper by Sourav Chatterjee and Persi Diaconis at Stanford on importance sampling in multivariate problems). That said, there is a class of high-dimensional problems where the Bootstrap filter can be integrated seamlessly, namely, state-space models.

2.1 A simple sequential Monte Carlo (SMC) algorithm

We shall now see an application of the Bootstrap filter to state-space models, a class of dynamic hidden-variable models. The following presentation is adapted from a seminal paper: *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, by Gordon, Salmond, Smith (1993) (henceforth GSS).

We begin with a general state-space model that consists of two parts: an observation equation of the form

$$x_t \sim f_t(\cdot \mid \theta_t, \phi), \qquad t = 1, \ldots, T, \tag{1}$$

and a state equation represented by a Markov process given by

$$\theta_t \sim q_t(\cdot \mid \theta_{t-1}, \psi). \tag{2}$$

Denoting $\mathbf{x}_t = (x_1, \ldots, x_t)'$, we focus on the posterior distribution $\pi(\theta_t \mid \mathbf{x}_t)$ of the unknown state θ_t given the observations up to time t. For simplicity of exposition, we assume that the starting distribution $\theta_0 \sim q_0$ and the parameters (ϕ, ψ) are known, although this assumption is not necessary for general SMC algorithms.

Example. Consider a Gaussian random walk model in which $x_t \mid \theta_t \sim N(\theta_t, V)$ and $\theta_t \mid \theta_{t-1} \sim N(\theta_{t-1}, W)$ for $t = 0, 1, \ldots, T = 100$. In this case, the posterior $p(\theta_t \mid \mathbf{x}_t)$ is Gaussian and analytic expressions for the posterior mean and variance are available via the standard Kalman filter, which enables us to evaluate the algorithm we develop next.

General expressions for the posterior distributions $\pi(\theta_t \mid \mathbf{x}_t)$ can be obtained recursively from Bayes theorem and the Markovian properties of the state space according to the following algorithm:

Algorithm : Bayesian filtering

Initialization: Set $\pi(\theta_0 \mid \mathbf{x}_0) := q_0(\theta_0)$.

Prediction: Using the Markovian property,

$$\pi(\theta_t \mid \mathbf{x}_{t-1}) = \int q_t(\theta_t \mid \theta_{t-1}) \pi(\theta_{t-1} \mid \mathbf{x}_{t-1}) d\theta_{t-1}. \tag{3}$$

[Since $\pi(\theta_t, \theta_{t-1} \mid \mathbf{x}_{t-1}) = q_t(\theta_t \mid \theta_{t-1}) \pi(\theta_{t-1} \mid \mathbf{x}_{t-1})$ by the Markovian property]

Update: Using Bayes rule and again exploiting the Markovian property,

$$\pi(\theta_t \mid \mathbf{x}_t) = \frac{f_t(x_t \mid \theta_t)\, \pi(\theta_t \mid \mathbf{x}_{t-1})}{m_t}, \tag{4}$$

where $m_t = \int f_t(x_t \mid \theta_t)\, \pi(\theta_t \mid \mathbf{x}_{t-1}) d\theta_t$.

For the Gaussian random walk model, the "Prediction" and "Update" steps can be carried out analytically.

In general, for non-conjugate models, the Bootstrap filter algorithm can be seamlessly adapted within the above recursive structure to make a transition from $\pi(\theta_{t-1} \mid \mathbf{x}_{t-1})$ to $\pi(\theta_t \mid \mathbf{x}_t)$ whenever it is possible to conveniently sample from the state equation $q_t(\cdot \mid \theta_{t-1})$ and to evaluate $f_t(\cdot \mid \theta_t)$. Specifically, suppose we are provided J independent samples $\{\theta_{t-1}^{(j)}\}_{j=1}^J$ from from $\pi(\theta_{t-1} \mid \mathbf{x}_{t-1})$. We run the following algorithm to approximate $\pi(\theta_t \mid \mathbf{x}_t)$.

Bootstrap Filtering in state-space models:

1. For $j = 1, \ldots, J$, draw $\theta_t^{(j)}$ from $q_t(\cdot \mid \theta_{t-1}^{(j)}, \psi)$. [for example, in the Gaussian example above, we draw $\theta_t^{(j)} \sim N(\theta_{t-1}^{(j)}, W)$]
2. Calculate weights
$$w_t^{(j)} = \frac{f_t(x_t \mid \theta_t^{(j)})}{\sum_{j'=1}^J f_t(x_t \mid \theta_t^{(j')})}$$

The algorithm outputs $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^J$, which provide a weighted approximation to the posterior. We can resample J times independently from this discrete distribution to obtain the independent samples needed to move from step t to $t+1$. Thus, it is straightforward to write a recursive algorithm to sample from $\pi_s(\theta_s \mid \mathbf{x}_s)$ for any $s \geq 1$. We initialize the algorithm with J prior samples from q_0 and recursively cycle through 3 steps: (i) sample, (ii) calculate weights, and (iii) resample.

In the normal example, I implemented the Bootstrap filtering algorithm. Figure 1 shows histograms of posterior samples obtained from the BF algorithm states θ_t with the true posterior density overlaid for time points $25, 50, 75$ and 100. (the true posterior can be exactly calculated recursively in the Gaussian random walk model as mentioned before; see HW problem!)

We next apply the Bootstrap filter to a non-linear state-space model previously considered in GSS.

$$\theta_t = 0.5\theta_{t-1} + 25\theta_{t-1}/(1+\theta_{t-1}^2) + 8\cos\{1.2(t-1)\} + w_t, \quad w_t \sim \mathrm{N}(0, W), \tag{5}$$
$$x_t = \theta_t^2/20 + v_t, \quad v_t \sim \mathrm{N}(0, V), \tag{6}$$

with $W = 10$, $V = 1$ and initial state $\theta_0 = 0.1$. The state evolution is highly non-linear and the unknown states θ_t enter the likelihood of the observed y_t through a quadratic transformation, rendering the posterior distributions of some of the states to be bimodal. As in [**?**], we first plot a 50 point realization from the state equation in Figure 2. Gordon et al., (1993) implemented their bootstrap filter algorithm initialized with $J = 500$ prior draws from the prior $p(x_0) \equiv \mathrm{N}(0, 2)$ and proceeding to resample 500 points from the discrete approximation to the posterior $p(\theta_t \mid \mathbf{x}_t)$ at each t. Their procedure comprehensively beat the ensemble Kalman filter (EKF) in recovering the posterior means for θ_j and provided significantly tighter 95% credible intervals. An implementation of the bootstrap filter algorithm is shown in Figure 3; I used the same legends and axis limits as in Figure 3 of Gordon et al., (1993) for ease of visual comparison.

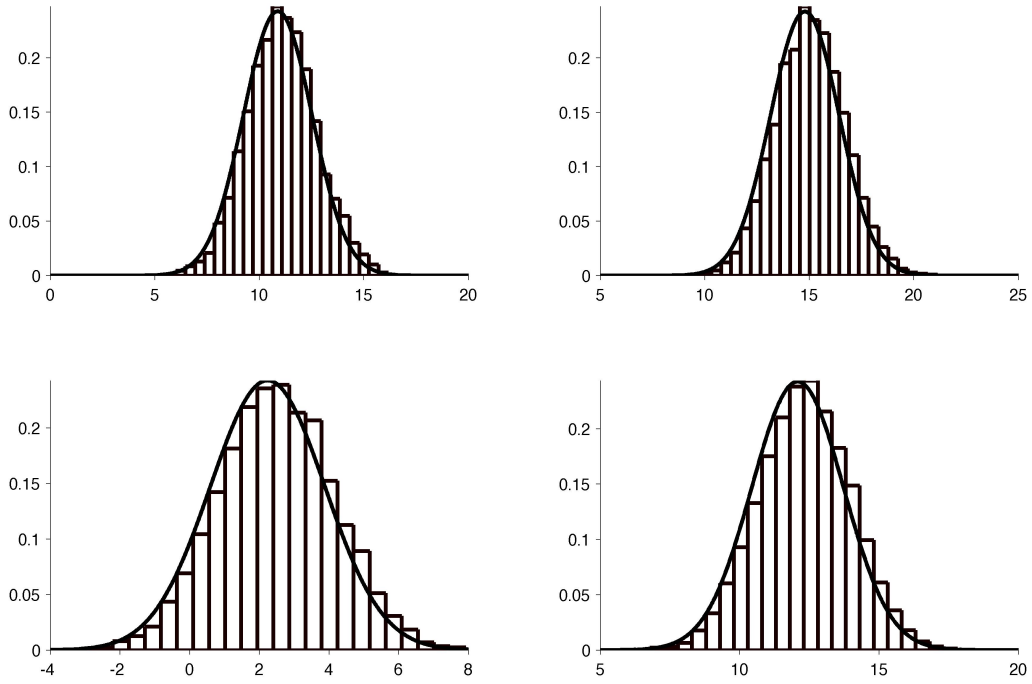

Figure 1: State estimation in random walk model. Histograms of posterior samples from Bootstrap filter for states θ_t for $t = 25, 50, 75, 100$ with the true posterior density overlaid.

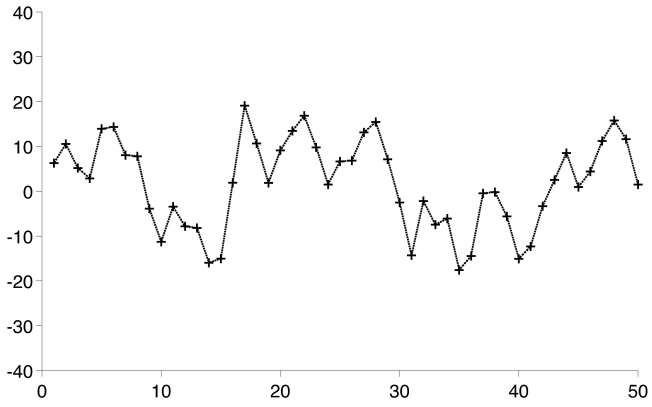

Figure 2: 50 point realization from (5) with initial state $\theta_0 = 0.1$.

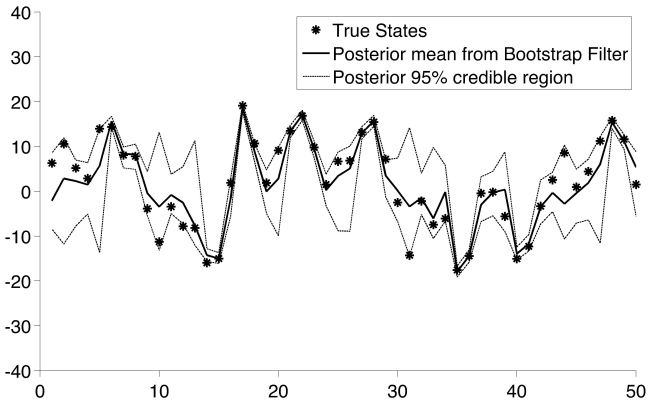

Figure 3: Bootstrap filter estimate of posterior mean and 95% credible regions for θ_t.