Nonlinear state esimation method: particle filtering

Rachel Han

December 4th, 2019

1 Introduction

Nonlinear state estimation refers to estimating the true *state* of a dynamic model where the underlying relationship between the *observations* $(X_n)_{n\geq 1}$ (which are often corrupted versions of the *state*) and the *state* $(Y_n)_{n\geq 1}$ where the relationship between X_n and Y_n is nonlinear. There are a handful of examples listed which [11] lists out: the states $(X_n)_{n\geq 1}$ could be unobserved signals in signal processing [10], some characteristic of shapes observed from sequence of images [1] and velocity of an object in a multi-target tracking problem in [7]. These type of models where the stochastic process we are interested in is 'hidden' and only inferable through some observations are called Hidden Markov Models. This term will be defined in the next section.

We are using the notation $x_{1:n}$ for $\{x_1, x_2, ..., x_n\}$, where $x_{1:n} \in \mathcal{X}^n$. For a filtering problem, one is interested in approximating the probability density of current state conditioned on the previous observations, $p(x_t|y_1, y_2, ..., y_t)$ at some t. Kalman filter is known to have a closed-form solution to this problem given that all noise is linear and Gaussian. For nonlinear state estimation problems, one can use the extended Kalman filter which linearizes about the unknown state. However, estimated the probability density function is still approximated by a Gaussian [7]. To better approximate the underlying distribution, particle filtering (or sequential Monte Carlo) methods can be used to approximate the conditional probability density $p(x_t|y_1, y_2, ..., y_t)$. The basis of Sequential Monte Carlo (SMC) is sequential importance sampling (SIS) and resampling. SIS and resampling methods will be discussed further in later sections. First, we define Hidden Markov Models.

1.1 Hidden Markov Model

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. We adot the definition of Markovian process from [3],

Definition 1. Suppose we have a stochastic process $X = (X_t)_{t\in\mathbb{T}}$ and is adapted to filtration $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{T}}$. X is *Markovian with relative to* $\mathcal{F}$ if for every time t, the past $\mathcal{F}_t$ and $\mathcal{G}_\infty^t = \sigma\{X_u : u \geq t, u \in \mathbb{T}\}$ are conditionally independent given the present state X_t.

The definition says that if a process X is Markovian, the future events are only determined by the present state X_t and is independent of the past. We can define *Hidden Markov Model (HMM)* as follows,

Definition 2. Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be discrete time stochastic processes. (X_n, Y_n) is a HMM if

- X_n is a Markov process that is not directly observable with a initial distribution μ that describes the initial state, and the transitional kernel $K(x, B) = \mathbb{P}(X_n \in B|X_1 = x_1, ..., X_{n-1} = x_{n-1}) = \mathbb{P}(X_n \in B|X_{n-1} = x_{n-1})$, for $B \in \mathcal{B}(\mathbb{R}^d)$. Let us call the associated density function $f(x_n|x_{n-1})$.
- Y_n are observations whose marginal distribution is given by $\mathbb{P}(Y_n \in A|X_1 = x_1, ..., X_n = x_n) = \mathbb{P}(Y_n \in A|X_n = x_n)$, for any measurable set $A \in \mathcal{B}(\mathbb{R}^d)$. Let us denote the marginal density as $g(y_n|x_n)$.

Therefore, HMM can be completely described by the initial measure μ, the transition probability kernel density f and the marginal density g.

In the light of Baye's Theorem, according to [16] the posterior distribution is given by

$$p(x_{1:n}|y_{1:n}) = \frac{p(y_{1:n}|x_{1:n})p(x_{1:n})}{p(y_{1:n})}. \tag{1.1}$$

We can recursively define the evolving posterior as

$$p(x_{1:n+1}|y_{1:n+1}) = p(x_{1:n}|y_{1:n})\frac{p(y_{n+1}|x_{n+1})p(x_{n+1}|x_{1:n})}{p(y_{n+1}|y_{1:n})} \tag{1.2}$$

This allows us to update the posterior and predict the next state

$$p(x_n|y_{1:n-1}) = \int p(x_n|x_{n-1})p(x_{n-1}|y_{1:n-1})dx_{n-1} \qquad \text{Predict} \tag{1.3}$$

$$p(x_n|y_{1:n}) = \frac{p(y_n|x_n)p(x_n|y_{1:n-1})}{\int p(y_n|x_n)p(x_n|y_{1:n-1})} \qquad \text{Update} \tag{1.4}$$

In general, these conditional distributions involving high dimensional integrals is hard to compute. This is why methods like Monte Carlo exist, because it allow us to approximate these quantities without having to compute the integrals, recursively. Additionally, this approximation can be done in done sequentially. This forms probabilistic dynamic system as defined in [11]:

Definition 3. A sequence of evolving probability distributions $\{\pi_n(x_{1:n})\}_{n\geq 1}$, where n represents the n-th time step is called a probabilistic dynamic system.

Sequential Monte Carlo methods is a device to evolve the probabilistic dynamic system. For the rest of the paper, we denote the posterior $p(x_{1:n}|y_{1:n})$ as a general probability density of interest $\pi_n(x_{1:n})$, known up to a normalizing constant.

$$\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n}, \quad Z_n = \int \gamma_n(x_{1:n})dx_{1:n}. \tag{1.5}$$

2 Perfect Sampling

In Monte Carlo methods, we can use empirical measures to approximate the desired distribution, $\pi_n(x_{1:n})$ where n represents the dimension of the probability space $\mathcal{X}^n$ [4]. Suppose we can sample N independent and identically distributed multivariate random variables (or particles) $\{X^i_{1:n}; i = 1, ..., N\}$ from the density $\pi_n(x_{1:n})$, for a fixed n. Given these sampled random variables, we are interested in estimating the underlying density. We can do this by using the discrete empirical measure

$$\hat{\pi}_n(x_{1:n}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^i_{1:n}}(x_{1:n}) \tag{2.1}$$

Sampling directly from $\pi_n(x_{1:n})$ is called the perfect sampling. The advantage of perfect sampling is that the variance of estimating expectation for any test function φ_n, $\text{Var}(\mathbb{E}[\varphi_n(x)])$, decreases as we increase the number of samples (N) regardless of the dimension of our probability space $\mathcal{X}^n$ [4]. The variance expression is given as the following:

$$\text{Var}(\mathbb{E}[\varphi_n(x)]) = \frac{1}{N}\int \varphi_n^2(x_{1:n})\pi_n(x_{1:n})dx_{1:n} - \mathbb{E}[\varphi_n(x_{1:n})]^2. \tag{2.2}$$

We can see that it decreases at a rate of $\mathcal{O}(\frac{1}{N})$. However, sampling directly from the target distribution is not feasible since it may be complex-high dimensional. In general, generating random samples from non-standard distributions is not trivial. Additionally, our aim is to evolve the density function to accommodate the incoming data which increases the dimension of our problem each time step. Even if we were able to sample directly from the target distribution, the problem gets harder as the dimension increases [4]. Therefore, we explore the following methods.

3 Importance Sampling

Since $\pi_n(x_{1:n})$ is difficult to sample from, one can attempt to sample from a more tractable density (easier to sample from) which is close to $\pi_n(x_{1:n})$. This is called importance sampling. We call this substitute density the importance density, denoted by $q_n(x_{1:n})$. In this paper, we outline the key ideas of important sampling explained in Doucet et al [4]. We require that if $\pi_n(x_{1:n}) > 0$, then $q_n(x_{1:n}) > 0$. Importance sampling uses a simple identity. For brevity, we denote the random vector $x_{1:n}$ as $\boldsymbol{x}$. Let $\varphi(\boldsymbol{x})$ be any measurable function. Then for any test function φ_n,

$$\mathbb{E}_{\pi_n}[\varphi(\boldsymbol{x})] = \int_E \varphi(\boldsymbol{x})\pi_n(\boldsymbol{x})d\boldsymbol{x} = \left(\int_E \varphi(\boldsymbol{x})\frac{\gamma(\boldsymbol{x})}{q_n(\boldsymbol{x})}q_n(\boldsymbol{x})d\boldsymbol{x}\right)/Z_n = \mathbb{E}_{q_n}[w_n(\boldsymbol{x})\varphi(\boldsymbol{x})]/Z_n, \tag{3.1}$$

where w_n is the unnormalised weight function and Z_n is a normalising constant,

$$w_n(\boldsymbol{x}) = \frac{\gamma_n(\boldsymbol{x})}{q_n(\boldsymbol{x})}, \tag{3.2}$$

$$Z_n = \int w_n(\boldsymbol{x})q_n(\boldsymbol{x})d\boldsymbol{x}. \tag{3.3}$$

Therefore, the target distribution is given by

$$\pi_n(\boldsymbol{x}_{1:n}) = \frac{w_n(\boldsymbol{x})q_n(\boldsymbol{x})}{Z_n}. \tag{3.4}$$

Assuming that $q_n(\boldsymbol{x})d\boldsymbol{x}$ dominates $\pi_n(\boldsymbol{x})d\boldsymbol{x}$ and these measures are sigma finite, we can view $w(\boldsymbol{x})$ as the Radon-Nikodym derivative. Importance sampling draws N samples $X^i_{1:n} \sim q_n(x_{1:n})$. We can write down the Monte Carlo approximation of $q_n(x_{1:n})$ using the empirical measure:

$$\hat{q}_n(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^i_{1:n}}(\boldsymbol{x}). \tag{3.5}$$

Then the approximation of the normalising constant Z_n is

$$\hat{Z}_n = \int_E w_n(\boldsymbol{x})q_n(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} w_n(\boldsymbol{x}) \tag{3.6}$$

and the target distribution π_n is approximated by

$$\hat{\pi}_n(\boldsymbol{x}) = \frac{w(X^i_{1:n})\delta_{X^i_{1:n}}(\boldsymbol{x})}{\sum_{i=1}^{N} w_n(X^i_{1:n})}. \tag{3.7}$$

Given these approximation of distributions, we can compute approximate expectations of any test function φ_n. It is an exercise to show that estimate of the expectation of a random vector $\varphi_n(\boldsymbol{x})$ is

$$\mathbb{E}_{\hat{\pi}_n}[\varphi(\boldsymbol{x})] = \sum_{i=1}^{N} W^i_n\varphi_n(\boldsymbol{X}^i) \tag{3.8}$$

, where

$$W^i_n = \frac{w(X^i_{1:n})}{\sum_{i=1}^{N} w_n(X^i_{1:n})}. \tag{3.9}$$

It is important to check whether this estimate of expectation converges to the correct target $\mathbb{E}_{\pi_n}[\varphi_n(\boldsymbol{X}^i)]$ as $N \to \infty$. This is true indeed by law of large numbers. Furthermore, the asymptotic variance of importance sampling estimator can be given by the central limit theorem, see Theorem 1 below. A short proof of this theorem is given in [6]. In this paper, with the help of [14] we expand this proof. We state useful theorems that give the central limit theorem for importance sampling.

Theorem 1 (Levy's Continuity Theorem)**.** This theorem can be found in [5]. Let $\boldsymbol{X}_n = (X_1, X_2, ..., X_k)$, and $(\boldsymbol{X})_{n\geq 1}$ be sequence of random vectors. Let $(\boldsymbol{\phi}_n(\boldsymbol{t}))_{n\geq 1}$ be the corresponding sequence of characteristic functions, where $\boldsymbol{\phi}_n(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}^T\boldsymbol{X}_n}]$, for every $\boldsymbol{t} \in \mathbb{R}^k, n \in \mathbb{N}$. If $\phi_n(t)$ converges pointwise to its limit ϕ: $\phi_n(t) \to \phi(t)$ then the following are equivalent:

- $\boldsymbol{X}_n$ converges in distribution to some variable $\boldsymbol{X}$.
- $(\boldsymbol{X}_n)_{n\geq 1}$ is tight.
- $\boldsymbol{\phi}(\boldsymbol{t})$ is a characteristic function of $\boldsymbol{X}$.
- $\boldsymbol{\phi}(\boldsymbol{t})$ is a continuous function of $\boldsymbol{t}$, and is continuous at $t = 0$.

The proof for Levy's Continuity Theorem is omitted but can be found in [5].

Theorem 2 (Cramer Wold Theorem)**.** Let $\boldsymbol{X}_n = (X_{n1}, X_{n2}, ..., X_{nk})$ and $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ be random vectors. If $\boldsymbol{t}^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T\bar{X}$ for each $\boldsymbol{t} = (t_1, ..., t_k) \in \mathbb{R}^k$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

The proof is left as an exercise.

Theorem 3 (Multivariate Central Limit Theorem)**.** Let $\boldsymbol{X}_n = (X_1, X_2, ..., X_k)$ be a random vector in $\mathbb{R}^k$, and each $\boldsymbol{X}_n$ be independent and identically distributed for all $n \in \{1, ..., N\}$ with mean $\boldsymbol{\mu}$ and covariance Σ. Let $\bar{\boldsymbol{X}}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i$. Then

$$\sqrt{N}(\bar{\boldsymbol{X}}_N - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}_k(0, \Sigma), \tag{3.10}$$

where $\mathcal{N}_k$ is the $k-$ dimensional multivariate normal distribution.

Proof. Let $\boldsymbol{Z}_n = \boldsymbol{t}^T(\boldsymbol{X_n} - \boldsymbol{\mu})$ be a random scalar variable, and $(\boldsymbol{Z}_n)_{n\geq 1}$ be its sequence. Each $\boldsymbol{X}_n$ is independent and identically distributed with mean $\boldsymbol{\mu}$ and variance Σ. Then its affine transformation $\boldsymbol{Z}_n$ is independent and identically distributed with mean $\boldsymbol{0}$ and variance $\boldsymbol{t}^T\Sigma\boldsymbol{t}$. The mean of $\boldsymbol{Z_n}$ is obtained by $\mathbb{E}[\boldsymbol{t}^T(\boldsymbol{X_n}-\boldsymbol{\mu})] = \sum_{i=1}^N t_i(\mathbb{E}[X_{ni}-\mu_i])] = 0$. The variance is $\mathrm{Var}(\sum_{i=1}^N t_i(X_{ni}-\mu_i)) = \sum_{i=1}^N t_i^2\mathrm{Var}(X_{ni}-\mu_i) = \boldsymbol{t}^T\Sigma\boldsymbol{t}$. Note that the covariance terms are 0 because each X_n is independent. Therefore, by the univariate central limit theorem $\sqrt{N}(Z_n) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{t}^T\Sigma\boldsymbol{t})$. Now we can apply Cramer-Wold theorem directly. Note that by the property affine transformation of multivariate normal, $\mathcal{N}(0, \boldsymbol{t}^T\Sigma\boldsymbol{t}) = \boldsymbol{t}^T \cdot \mathcal{N}_k(0, \Sigma)$. Therefore, since $\boldsymbol{t}^T(\boldsymbol{X_n} - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{t}^T \cdot \mathcal{N}_k(0, \Sigma)$, $\sqrt{N}(\bar{\boldsymbol{X}}_N - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}_k(0, \Sigma)$. □

Theorem 4 (Delta Method)**.** From [15], we state the univariate delta method. Let Y_n be a sequence of random variables. If for some θ

$$\sqrt{n}(Y_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{3.11}$$

for some differentiable g such that $g'(\theta) \neq 0$, then

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{D} \mathcal{N}(0, \sigma^2(g'(\theta))^2). \tag{3.12}$$

Proof. We reiterate the proof from [15]. Taylor expanding $g(Y_n)$ about θ, we get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)(Y_n - \theta)^2}{2} + \frac{g'''(\theta)(Y_n - \theta)^3}{6} + ...$$

Let $R = \frac{g''(\theta)(Y_n-\theta)^2}{2} + \frac{g'''(\theta)(Y_n-\theta)^3}{6} + ...$ Rearranging the terms and multiplying by $\sqrt{n}$,

$$\sqrt{n}(g(Y_n) - g(\theta)) = \sqrt{n}[g'(\theta)(Y_n - \theta) + R] \tag{3.13}$$

We invoke Slutsky's theorem that we won't prove but state. The theorem states that for sequence of random variables W_n and Z_n, if $W_n \xrightarrow{D} W$ and $Z_n \xrightarrow{prob} c$ where c is a deterministic scalar, $W_n + Z_n \xrightarrow{D} W + c$ and $W_nZ_n \xrightarrow{D} cW$. We know by assumption, $Y_n \xrightarrow{prob} \theta$, so $\frac{g''(\theta)(Y_n-\theta)^2}{2} + \frac{g'''(\theta)(Y_n-\theta)^3}{6} + ... \xrightarrow{prob} 0$.

Also $\sqrt{n}g'(\theta)(Y_n - \theta) \xrightarrow{D} \mathcal{N}(0, g'(\theta)^2\sigma^2)$. This is left as an exercise to show. Then by Slutsky's theorem, $\sqrt{n}[g'(\theta)(Y_n - \theta) + R] \xrightarrow{D} \mathcal{N}(0, g'(\theta)^2\sigma^2) + 0$, so by (3.13),

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{D} \mathcal{N}(0, g'(\theta)^2\sigma^2).$$

□

Proposition 1. *Central limit theorem for importance sampling.* Suppose $(\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N)$ are sequence of i.i.d random vectors. Then the following holds,

$$\sqrt{N}\left(\sum_{i=1}^{N} W^i \varphi(\boldsymbol{X}^i) - \mathbb{E}_\gamma[\varphi(\boldsymbol{x})]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{3.14}$$

where

$$\sigma^2 = \int \frac{\gamma(\boldsymbol{x})^2}{q(\boldsymbol{x})}(\varphi(\boldsymbol{x}) - \mathbb{E}_\gamma(\varphi))^2. \tag{3.15}$$

We suppress the subscript n to denote the dimension, because this proposition holds for any fixed n.

Proof. Let Y_n be a vector

$$Y_N = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{X}^i)\varphi(\boldsymbol{X}^i) \\ \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{X}^i) \end{bmatrix}. \tag{3.16}$$

Firstly, by the multidimensional Central Limit Theorem,

$$\sqrt{N}(Y_N - \mathbb{E}_q[Y_N]) \to \mathcal{N}(0, \Sigma) \tag{3.17}$$

as $N \to \infty$ where

$$\Sigma = \begin{bmatrix} \mathrm{Var}(Y_{N,1}) & \mathrm{Cov}(Y_{N,1}, Y_{N,2}) \\ \mathrm{Cov}(Y_{N,2}, Y_{N,1}) & \mathrm{Var}(Y_{N,1}). \end{bmatrix}. \tag{3.18}$$

The expectation of Y_N can be computed as follows:

$$\mathbb{E}_q[Y_N] = \begin{bmatrix} \mathbb{E}_q[Y_{N,1}] & \mathbb{E}_q[Y_{N,2}]^T \end{bmatrix}. \tag{3.19}$$

$$\mathbb{E}_q[Y_{N,1}] = \mathbb{E}_q[\frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{X}^i)\varphi(\boldsymbol{X}^i)] = \frac{1}{N}\mathbb{E}_q\left[\sum_{i=1}^{N} \frac{\gamma(\boldsymbol{X}^i)}{q(\boldsymbol{X}^i)}\varphi(\boldsymbol{X}^i)\right] \tag{3.20}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_q\left[\frac{\gamma(\boldsymbol{X}^i)}{q(\boldsymbol{X}^i)}\varphi(\boldsymbol{X}^i)\right] = \frac{1}{N}\sum_{i=1}^{N}\int \frac{\gamma(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}\varphi(\boldsymbol{x}_i)q(\boldsymbol{x}_i)d\boldsymbol{x} \tag{3.21}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \gamma(\boldsymbol{x}_i)\varphi(\boldsymbol{x}^i)d\boldsymbol{x} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_\gamma[\varphi(\boldsymbol{X}^i)] \xrightarrow{a.s} \mathbb{E}_\gamma(\varphi(\boldsymbol{x})) \tag{3.22}$$

In (3.21), we care able to exchange the limit and sum because of the finiteness. (3.22) is by convergence of empirical distribution, $\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_\gamma[\varphi(\boldsymbol{X}^i)] \xrightarrow{a.s} \mathbb{E}_\gamma(\varphi(\boldsymbol{x}))$. $\mathbb{E}_q[Y_{N,2}]$ can be computed similarly as

$$\mathbb{E}_q[Y_{N,2}] = \frac{1}{N}\mathbb{E}[\sum_{i=1}^{N} w(\boldsymbol{X}^i)] \xrightarrow{a.s} \int \gamma(\boldsymbol{x})d\boldsymbol{x} = 1. \tag{3.23}$$

Therefore,

$$\mathbb{E}_q[Y_N] \xrightarrow{a.s} [\mathbb{E}_\gamma[\varphi(x)], 1] \tag{3.24}$$

Let $g(x,y) = \frac{x}{y}$ and $\mu = \mathbb{E}[\varphi(x)]$. Then $g(Y_N) = \sum_{i=1}^N W_n^i \varphi_n(X^i)$ and $g([\mu,1]) = \mu$. The gradient $\nabla g = \left(1/y, -x/y^2\right)$. Therefore, by the delta method,

$$\left(\sum_{i=1}^N W_n^i \varphi_n(X^i) - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla g([\mu,1])^T \cdot \Sigma \cdot \nabla g([\mu,1])\right). \tag{3.25}$$

where $\nabla g([\mu,1])^T \cdot \Sigma \cdot \nabla g([\mu,1])$ can be expanded as:

$$\nabla g([\mu,1])^T \cdot \Sigma \cdot \nabla g([\mu,1]) = \mathrm{Var}(Y_{N,1}) - 2\mu \mathrm{Cov}(Y_{N,1}, Y_{N,2}) + \mu^2 \mathrm{Var}(Y_{N,2}) \tag{3.26}$$

By convergence of empirical distribution, as $N \to \infty$,

$$\nabla g([\mu,1])^T \cdot \Sigma \cdot \nabla g([\mu,1]) \xrightarrow{a.s} \frac{1}{N}[\mathrm{Var}(w(x)\varphi(x)) - 2\mu \mathrm{Cov}(w(x)\varphi(x), w(x)) + \mu^2 \mathrm{Var}(w(x))] \tag{3.27}$$

$$= \frac{1}{N}\mathrm{Var}(\mu w(x) - w(x)\varphi(x)) \qquad \text{By property of variance of two r.v.} \tag{3.28}$$

$$= \frac{1}{N}\mathrm{Var}((\mu - \varphi(x))w(x)) \tag{3.29}$$

$$= \frac{1}{N}[\mathbb{E}[w^2(x)(\mu - \varphi(x))^2] - \mathbb{E}[w(x)(\mu - \varphi(x))]^2] \tag{3.30}$$

$$= \frac{1}{N}\left(\int \frac{\gamma^2(x)}{q^2(x)}(\mu - \varphi(x))^2 q(x)dx - \left(\int \frac{\gamma(x)}{q(x)}(\mu - \varphi(x))q(x)dx\right)^2\right) \tag{3.31}$$

$$= \frac{1}{N}\left(\int \frac{\gamma^2(x)}{q(x)}(\mu - \varphi(x))^2 dx - (\mathbb{E}_\gamma[\mathbb{E}_\gamma[\varphi(x)]] - \mathbb{E}_\gamma[\varphi(x)])^2\right) \tag{3.32}$$

$$= \frac{1}{N}\int \frac{\gamma^2(x)}{q(x)}(\mathbb{E}_\gamma(\varphi(x)) - \varphi(x))^2 dx. \tag{3.33}$$

Let $\sigma^2 = \int \frac{\gamma^2(x)}{q(x)}(\mathbb{E}_\gamma(\varphi(x)) - \varphi(x))^2 dx$.. Then,

$$\left(\sum_{i=1}^N W_n^i \varphi_n(X^i) - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right). \tag{3.34}$$

□

This finishes the discussion of asymptotic variance of $\mathbb{E}_{q_n}[\varphi_n(x)]$. Sometimes, the variance of the importance weights is of interest. The test function of interest may change every time step [4]. For example, in a sequential setting, we are interested in $\varphi_{n-1}(x_{1:n-1})$ at a time step and at the next time step, $\varphi_n(x_{1:n})$. Choosing q such to minimize variance of $\mathbb{E}_{q_n}[\varphi_{n-1}(x_{1:n-1})]$ is a bad idea. Instead, the variance of the importance weights w_n independent of the test function. The expression of the variance is given below.

Proposition 2. The variance of the weight function $w(\boldsymbol{x})$ is given by

$$\int \frac{\gamma_n^2(\boldsymbol{x})}{q_n(\boldsymbol{x})} d\boldsymbol{x} - \mathbb{E}[\gamma_n(x)]^2 \tag{3.35}$$

Proof. It follows from the definition of variance,

$$\mathbb{V}(w(x)) = \mathbb{E}_{q_n}[w^2(x)] - \mathbb{E}_{q_n}[w(x)]^2 \tag{3.36}$$

$$= \int \left(\frac{\gamma_n(\boldsymbol{x})}{q_n(\boldsymbol{x})}\right)^2 q_n(\boldsymbol{x})dx - \left(\int \frac{\gamma_n(\boldsymbol{x})}{q_n(\boldsymbol{x})} q_n(\boldsymbol{x}) d\boldsymbol{x}\right)^2 \tag{3.37}$$

$$= \int \frac{\gamma_n^2(\boldsymbol{x})}{q_n(\boldsymbol{x})} d\boldsymbol{x} - \mathbb{E}[\gamma_n(x)]^2. \tag{3.38}$$

□

Therefore, it can be checked that the relative variance of the normalising constant is given by

$$\frac{\mathrm{Var}(\hat{Z}_n)}{Z_n^2} = \frac{1}{N}\left(\int \frac{\pi_n^2(\boldsymbol{x})}{q_n(\boldsymbol{x})} d\boldsymbol{x} - 1\right). \tag{3.39}$$

Observe that (3.39) is minimized when $q_n = \pi_n$. Of course, making such a choice defeats the purpose of IS. Often the difficulty of this method is choosing q_n such that it is close to as γ_n as possible.

Example 1. Here is a very simple example adapted from [13] to illustrate the idea of importance sampling. Consider the function

$$I(a) = \int_a^\infty e^{-x} dx \tag{3.40}$$

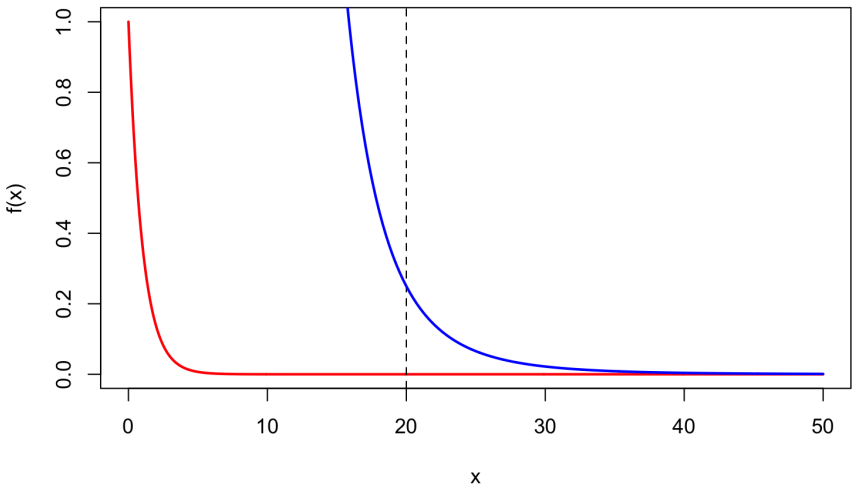
$f(x) = e^{-x}$ is our target density decays rapidly. We wish to compute $I(20)$. The importance sampling formulation is as follows:

$$\bar{I}(a) = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X^i)}{q(X^i)} \tag{3.41}$$

for samples $X^i, i = 1, ..., N$ drawn from q. We choose a importance density q such that it has a heavier tail at $x = 20$ but with a similar shape. This will allow us to sample easily from the "important" location, which is around $x = 20$. Let q be the pareto probability density function,

$$q(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}. \tag{3.42}$$

We set the parameters to $x_m = 20$ and $\alpha = 5$. Next, we need to sample N random samples from the pareto probability distribution function. We can do this by the inverse transform sampling. The samples are given



by

$$X = \frac{x_m}{U^{\frac{1}{\alpha}}} \sim \mathrm{Pareto}(x_m, \alpha) \tag{3.43}$$
$$U \sim \mathcal{U}(0,1),$$

where $\mathcal{U}$ is the uniform distribution. Given N samples X^i, we can compute $\bar{I(a)}$ as in Equation (3.41). The result obtained is $\bar{I}(a) = 2.061185e-09$ and $I(a) = 2.061154e-09$. We can see that $\bar{I(a)}$ approximates $I(a)$ quite well.

Importance sampling alleviates the difficulty of directly sampling from an intractable distribution π_n by choosing a better distribution to sample from. However, note that it does not solve the problem of increasing dimensionality.

4 Sequential Importance Sampling

The complexity of importance sampling depends on the dimension n. Therefore, as the distribution evolves, the complexity of sampling from distribution will grow as a function of n. Sequential importance sampling (SIS) alleviates this difficulty by fixing the computational complexity at each time step. Many Monte Carlo problems arise where the target distribution π_n has a decomposable structure [16].

$$\pi_n(x_{1:n}) = \pi_{n-1}(x_{1:n-1})\pi_n(x_n|x_{1:n-1}). \tag{4.1}$$

In SIS, we assume a similar structure for importance density q_n:

$$q_n(x_{1:n}) = q_{n-1}(x_{1:n-1})q_n(x_n|x_{1:n-1}) \tag{4.2}$$

$$= q_1(x_1)\prod_{k=2}^{n} q_k(x_k|x_{1:k-1}). \tag{4.3}$$

In SIS, we first draw a sample X_1^i from $q_1(x_1)$ at time 1. Then, at time 2, we would draw the second sample X_2^i from $q_2(x_2|X_1^i, X_2^i)$. At k-th time, one would draw X_k^i from $q_k(x_k|X_{1:k-1}^i)$. We would sample in this way for $i = 1, ..., N$. The unnormalised weights w_n can be written recursively,

$$w_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{q_n(x_{1:n})} \tag{4.4}$$

$$= \frac{\gamma_{n-1}(x_{1:n-1})}{q_{n-1}(x_{1:n-1})}\frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})q_n(x_n|x_{1:n-1})} \tag{4.5}$$

$$= w_{n-1}(x_{1:n-1})\frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})q_n(x_n|x_{1:n-1})} \tag{4.6}$$

Again, the variance of w_n depends on the choice of $q_n(x_n|x_{1:n-1})$. From (3.39), the optimal importance density is given by $q_n^{opt}(x_n|x_{1:n-1}) = \pi_n(x_n|x_{1:n-1}) = \gamma_n(x_n|x_{1:n-1})/\int \gamma_n(x_n|x_{1:n-1})dx_n$. Then,

$$\alpha_n^{opt} = \frac{\gamma_n(x_{1:n-1})}{\gamma_{n-1}(x_{1:n-1})}. \tag{4.7}$$

In most cases this optimality cannot be achieved since we do not know π_n. In the context of particle filtering, we are interested in the joint prior $p(x_{1:n}|y_{1:n})$. For simplicity, we the case consider where $\pi_n(x_{1:n}) = \gamma_n(x_{1:n}) = p(x_{1:n}|y_{1:n})$. The observations up to time n, $y_{1:n}$ determine the distribution at the n-th time. Therefore $\gamma_n(x_{1:n-1})$ correspond to $p(x_{1:n-1}|y_{1:n})$. Therefore, the optimal importance weight at time $n+1$ is given by

$$w_{n+1} = w_n\frac{p(x_{1:n}|y_{1:n+1})}{p(x_{1:n}|y_{1:n})}. \tag{4.8}$$

Kong et al. [9] showed that the sequential importance weights w_n are martingales with respect to n in the context of sequential imputation. Here we list out the details of this theorem in [9].

Theorem 5 (Kong-Liu-Wong Theorem [9])**.** The sequential importance weights w_n is a martingale sequence in n with random variables x and y. This implies that the variance of the weights are increasing with respect to n.

Proof. Let $\mathcal{F}_n$ be a filtration over the probability space $(\Omega, \mathcal{H}, \mathbb{P})$: $\mathcal{F}_n = \sigma\{x_1, x_2, ..., x_n, y_1, y_2, ..., y_n\}$. Then it is clear that for each $n \in \mathbb{N}$, $w_n \in \mathcal{F}_n$, so $w)n$ is adapted to $\mathcal{F}_n$. Furthermore, we assume that $\mathbb{E}[w_n]$ is finite. It remains to show that $\mathbb{E}[w_{n+1}|\mathcal{F}_n] = w_n$. Since $w_n \in \mathcal{F}_n$, by conditional determinism,

$$\mathbb{E}[w_{n+1}|\mathcal{F}_n] = \mathbb{E}\left[w_n\frac{p(x_{1:n}|\sigma(y_{1:n}), \sigma(y_{n+1}))}{p(x_{1:n}|\sigma(y_{1:n}))}|\mathcal{F}_n\right] = w_n\mathbb{E}\left[\frac{p(x_{1:n}|\sigma(y_{1:n}), \sigma(y_{n+1}))}{p(x_{1:n}|\sigma(y_{1:n}))}|\mathcal{F}_n\right]. \tag{4.9}$$

Note that the random variable of interest in (4.9) is y_{n+1}. Let $\mathcal{G} = \sigma(y_1, y_2, ..., y_n), \mathcal{G}_1 = \sigma(x_1, x_2, ..., x_n)$ and $\mathcal{G}_2 = \sigma(y_{n+1})$. Therefore, the conditional density is $p(y_{n+1}|\mathcal{G}, \mathcal{G}_1)$. However, y_{n+1} is conditionally

independent of $x_{1:n}$ given $y_{1:n}$. This is because given observations up to present time, an observation at a future time is assumed to be independent of the present states (Markov property). It is known that if for any σ-algebras $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2$,

$$\mathcal{F}_1 \perp\!\!\!\perp_{\mathcal{F}} \mathcal{F}_2 \quad \text{iff} \quad \mathbb{P}[H|\mathcal{F}, \mathcal{F}_1] = \mathbb{P}[H|\mathcal{F}] \tag{4.10}$$

for each $H \in \mathcal{F}_2$ for the respective probability measure $\mathbb{P}$. Therefore, the conditional density is $p(y_{n+1}|\mathcal{G})$. Then,

$$\mathbb{E}[w_{n+1}|\mathcal{F}_n] = w_n \mathbb{E}\left[\frac{p(x_{1:n}|\sigma(y_{1:n}), \sigma(y_{n+1}))}{p(x_{1:n}|\sigma(y_{1:n}))}|\mathcal{F}_n\right] \tag{4.11}$$

$$= w_n \int \frac{p(x_{1:n}|\sigma(y_{1:n}), \sigma(y_{n+1}))}{p(x_{1:n}|\sigma(y_{1:n}))} p(y_{n+1}|\sigma(y_{1:n}))dy_{n+1}. \tag{4.12}$$

Again by conditional independence, $p(x_{1:n}|\sigma(y_{1:n}), \sigma(y_{n+1})) = p(x_{1:n}|\sigma(y_{1:n}))$. Therefore, since the joint distribution is the product of the marginal and conditional, $p(x_{1:n}|\sigma(y_{1:n}))p(y_{n+1}|\sigma(y_{1:n})) = p(x_{1:n}, y_{n+1}|\sigma(y_{1:n}))$. Equation (4.12) is

$$\mathbb{E}[w_{n+1}|\mathcal{F}_n] = w_n \int \frac{p(x_{1:n}, y_{n+1}|\sigma(y_{1:n})}{p(x_{1:n}|\sigma(y_{1:n}))} dy_{n+1} \tag{4.13}$$

$$= w_n \int p(y_{n+1}|\sigma(x_{1:n}), \sigma(y_{1:n}))dy_{n+1} \qquad \text{Conditional is joint divided by marginal.} \tag{4.14}$$

$$= w_n \cdot 1 \qquad \text{Integrating conditional over all } y_{n+1} \text{ gives 1.} \tag{4.15}$$

We have shown that $\mathbb{E}[w_{n+1}|\mathcal{F}_n] = w_n$. It is left as an exercise to show that $\text{Var}(w_n) \leq \text{Var}(w_{n+1})$. □

Therefore, even with $q_n^o pt$, the variance of w_n increases with n. Looking at (3.39), we see that as the importance density q_n deviates from π_n, the weights w_n become smaller with the growing variance. Increasing variance will drive most of the weights to zero except a few that are close to 1. The phenomenon where the weights become increasingly skewed is called the degeneracy. Although SIS can fix the dimensionality problem by restricting itself to low dimensional conditional importance densities, it cannot avoid the degeneracy when n gets too large.

5 Resampling

The degeneracy problem can be avoided by resampling. The idea behind is that we want to make sure the particles that are likely to be sampled are kept and remove the ones with low weights. A simple resampling is called the multinomial resampling method. After sampling from the importance density q_n we get the approximation of the target density (3.7),

$$\hat{\pi}_n(x_{1:n}) = \sum_{i=1}^{N} W_n^i \delta_{X_{1:n}^i(x_{1:n})}.$$

The normalized weights W_n^i from (A.2) are associated with each particles $X_{1:n}^i$. Multinomial resampling [7] proceeds as follows:

1. Obtain $\hat{\pi}_n$ from the samples $(X_{1:n}^1, X_{1:n}^2, ..., X_{1:n}^N)$.
2. Sample N times with replacement from $\hat{\pi}_n$ to obtain $(X_{1:n}^{1*}, X_{1:n}^{2*}, ..., X_{1:n}^{N*})$. According to [4], this is equal to the following:
   - Generate a random number representing the number of occurrences of each particle N_n^i, where it is sampled $N_n^i \sim \mathcal{M}(N, w_{1:n}^1, w_{1:n}^2, ..., w_{1:n}^N)$. Sampling from multinomial distribution where there are N possible outcomes with probability $w_{1:n}^i$ can be done in the following way. We subdivide a unit interval $[0, 1]$ into N pieces. Each piece is of different length proportional to some weight w^i and has value N^i. We pick a random number from $\mathcal{U}(0, 1)$ which would correspond to a location on the strip. Then we return N^i corresponding to the strip.

Figure 1: A visualization of multinomial sampling reduced to uniform sampling

3. Next, we assign equal weights $1/N$ to the number of occurrences of each particle $(N_n^i)_{j=1,...,N}$.

This method works because by resampling from $\hat{\pi}_n$, we can approximate the target distribution π_n in the limit as $N \to \infty$ as shown by Smith and Gelfand (1992). However, as shown by Chopin [2], the local variance of estimating the expectation of a test function φ_n is greater when we perform resampling than SIS. There are other resampling methods that attempt to reduce the variance, such as the residual resampling method by Liu and Chen [10]. In general, it is useful to resample only when there is a degeneracy because of this local variance drawback. In literature, Effective Sample Size (ESS) criterion is used to determine the degeneracy. There are many forms of ESS one can use. We present one type of ESS from [10] and [9]. For a fixed n, ESS is

$$ESS = \frac{N}{1 + C^2(w)}, \tag{5.1}$$

where $C(w)$ is the coefficient of variation

$$C^2(w) = \frac{1}{N} \sum_{i=1}^{N} (NW^i - 1)^2. \tag{5.2}$$

$C^2(w)$ is an approximation of the variance of importance weights when N is large [10]. Large $C^2(w)$ indicates that the number of surviving samples with non-zero weights is small. Typically the threshold of ESS is set to $N/2$.

6 Sequential Monte Carlo

Finally, the sequential Monte Carlo (SMC) can be articulated. A basic SMC method is Sequential Importance Resampling (SIR) which alternates between SIS and resampling. The steps of the algorithm is as follows.

Algorithm 1 Sequential Monte Carlo [4]

Sample X_n^i from $q_1(x_1)$, for $i = 1, ..., N$.
Compute the weights $w_1(X_1^i)$ and compute the normalized weights W_1^i.
if ESS $< N/2$ **then**
    Resample $\{W_1^i, X_1^i\}$ to get $\{\bar{W}_1^i, \bar{X}_1^i\} \leftarrow \{\frac{1}{N}, X_1^{i*}\}$
else
    $\{\bar{W}_1^i, \bar{X}_1^i\} \leftarrow \{W_1^i, X_1^i\}$.
end if
for $n = 2, 3, ...$ **do**
    Sample X_n^i from $q_n(x_n | X_{1:n-1}^{i*})$, for $i = 1, ..., N$.
    $X_{1:n}^i \leftarrow (\bar{X}_{1:n-1}^i, X_n^i)$
    Compute the incremental weights $\alpha_n(X_{1:n}^i)$ and the normalized weights $W_n^i \propto \bar{W}_n^i \alpha_n(X_{1:n}^i)$.
    if ESS $< N/2$ **then**
        Resample $\{W_n^i, X_{1:n}^i\}$ to get $\{\bar{W}_n^i, \bar{X}_{1:n}^i\} \leftarrow \{\frac{1}{N}, X_{1:n}^{i*}\}$
    else
        $\{\bar{W}_n^i, \bar{X}_{1:n}^i\} \leftarrow \{W_n^i, X_{1:n}^i\}$.
    end if
end for

Therefore at any point in time n, we are able to get approximation of π_n as $\bar{\pi}_n(x_{1:n}) = \sum_{i=1}^{N} \bar{W}_n^i \delta_{\bar{X}_{1:n}^i}(x_{1:n})$.

7 Open questions and research directions

Now that we know the theoretical justification of SMC methods, we can use it for applications of interest. A particular area I am interested in is estimating battery characteristics. For example, there is an active area of battery research devoted to create efficient estimation methods for the state of charge(SoC). The amount of charge in a battery with certain capacity is called SoC, like the battery indicator one would see on a laptop or phone (0% to 100% scale). One might notice that the battery indicator reflects the true SoC for a new device but as the device gets old, the estimated SoC would be less accurate. SoC cannot be measured directly but can be inferred from other measurements such as voltage and current. In fact, SoC is a function of solid lithium ion concentrations, which we denote by c.

Mainly, there are two battery models: equivalent circuit model and electrochemical model. The electrochemical models that describe the dynamics within the battery, which are made up of complex, deterministic partial differential algebraic equations (PDAE) to model Li-ion concentration, flux, electrolyte concentration, temperature, etc. Equivalent circuit model consists of open circuit voltage, resistor and capacitance. To formulate a particle filter problem, one would need to create a state-space model. There are different proposed state-space models from both of these models, see [8] and [17]. One would need to consider the complexity of the model when selecting which dynamics to include and choosing the right parameters is not a trivial task either.

The biggest challenge appears to be selecting the importance density. In [17], the authors use Dirac distribution to satisfy the boundary conditions in their electrochemical model based particle filter approach. [12] uses equivalent circuit model and Gaussian importance density function. However, using Gaussian may not be exploiting the full benefits particle filter that is suited for nonlinear and non-Gaussian problems. Picking a importance density that is justifiable is a crucial step in reducing the variance of the method. Then, one needs to consider the appropriate sampling technique to draw samples from the importance density. One can also experiment with various resampling methods.

A Exercises

1. Show that estimate of the expectation of a random vector $\varphi_n(\boldsymbol{x})$ is

$$\mathbb{E}_{\hat{\pi}_n}[\varphi(\boldsymbol{x})] = \sum_{i=1}^{N} W_n^i \varphi_n(\boldsymbol{X}^i), \tag{A.1}$$

where

$$W_n^i = \frac{w(X_{1:n}^i)}{\sum_{i=1}^N w_n(X_{1:n}^i)}. \tag{A.2}$$

Solution:

$$\mathbb{E}_{\hat{\pi}_n}[\varphi(\boldsymbol{x})] = \int \varphi_n(\boldsymbol{x})\hat{\pi}_n(\boldsymbol{x})d\boldsymbol{x} = \int \varphi_n(\boldsymbol{x})\hat{\pi}_n(d\boldsymbol{x}) \tag{A.3}$$

$$= \int \frac{w(X_{1:n}^i)}{\sum_{i=1}^N w_n(X_{1:n}^i)}\varphi_n(\boldsymbol{x})\delta_{X_{1:n}^i}(\boldsymbol{dx}) \tag{A.4}$$

$$= \sum_{i=1}^{N} \frac{w(X_{1:n}^i)}{\sum_{i=1}^N w_n(X_{1:n}^i)}\varphi_n \circ X_{1:n}^i = \sum_{i=1}^{N} W_n^i \varphi_n(X_{1:n}^i). \tag{A.5}$$

2. Using Levy's Continuity Theorem, prove Cramer-Wold theorem:

Theorem 6 (Cramer Wold Theorem)**.** Let $\boldsymbol{X}_n = (X_{n1}, X_{n2}, ..., X_{nk})$ and $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ be random vectors. If $\boldsymbol{t}^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T\boldsymbol{X}$ for each $\boldsymbol{t} = (t_1, ..., t_k) \in \mathbb{R}^k$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

Solution:
Suppose. $\boldsymbol{t}^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T\bar{\boldsymbol{X}}$. Then, $\mathbb{E}[e^{i\boldsymbol{t}^T\boldsymbol{X}_n}] \xrightarrow{D} \mathbb{E}[e^{i\boldsymbol{t}^T\boldsymbol{X}}]$ so the characteristic functions converge. Then by Lev's continuity theorem, $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

3. Show that if $X_n = \sqrt{n}(Y_n - \theta)$ and $X \xrightarrow{D} \mathcal{N}(0, \sigma^2)$, then $\sqrt{n}g'(\theta)(Y_n - \theta) \xrightarrow{D} \mathcal{N}(0, g'(\theta)^2\sigma^2)$, as used in the proof of multivariate central limit theorem.
Solution:
Suppose $X_n \sim \mathcal{N}(0, \sigma^2)$ with variance σ^2. Let $Z_n = g'(\theta)X_n$. Then $\mathrm{Var}(Z_n) = \mathrm{Var}(g'(\theta)X_n) = g'(\theta)^2\mathrm{Var}(X_n) = g'(\theta)^2\sigma^2$. Since the characteristic function determines the distribution, using the fact that $\mathbb{E}[e^{itX_n}] = e^{\frac{1}{2}\sigma^2t^2}$, $\mathbb{E}[e^{itZ_n}] = \mathbb{E}[e^{itg'(\theta)X_n}] = \mathbb{E}[e^{i(tg'(\theta))X_n}] = e^{\frac{1}{2}(tg'(\theta))^2\sigma^2}$. This characteristic function of Z_n is equal to that of $\mathcal{N}(0, g'(\theta)^2\sigma^2)$. Therefore, $Z_n \sim \mathcal{N}(0, g'(\theta)^2\sigma^2)$. So if $X_n = \sqrt{n}(Y_n - g(\theta)) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ then $Z_n = g'(\theta)\sqrt{n}(Y_n - \theta) \xrightarrow{D} \mathcal{N}(0, g'(\theta)^2\sigma^2)$.

4. Show that $\mathrm{Var}(w_n) \leq \mathrm{Var}(w_{n+1})$ as a consequence of Kong-Liu-Wong Theorem.
Solution:
The theorem tells us that $\mathrm{Var}(w_n) = \mathrm{Var}(\mathbb{E}[w_{n+1}|\mathcal{F}_n])$. It remains to show that

$$\mathrm{Var}(\mathbb{E}[w_{n+1}|\mathcal{F}_n]) \leq \mathrm{Var}(w_{n+1}).$$

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be our probability space, and let V be a collection of all random variables with finite variance. V is a Hilbert space with under inner product $< X, Y >= \mathbb{E}[XY]$, for $X, Y \in V$ [18]. So $w_{n+1} \in V$. $\mathcal{F}_n$ which is sigma algebra generated by a collection random variables in V is a closed subspace of V. w_{n+1} The two variances of interest are given by:

$$\mathrm{Var}(w_{n+1}) = \mathbb{E}[(w_{n+1})^2] - \mathbb{E}[w_{n+1}]^2, \tag{A.6}$$

$$\mathrm{Var}(\mathbb{E}[w_{n+1}|\mathcal{F}_n]) = \mathbb{E}[\mathbb{E}[w_{n+1}|\mathcal{F}_n]^2] - \mathbb{E}[\mathbb{E}[w_{n+1}|\mathcal{F}_n]]^2 = \mathbb{E}[\mathbb{E}[w_{n+1}|\mathcal{F}_n]^2] - E[w_{n+1}]^2. \tag{A.7}$$

It remains to show that

$$\mathbb{E}[\mathbb{E}[w_{n+1}|\mathcal{F}_n]^2] \leq \mathbb{E}[(w_{n+1})^2]. \tag{A.8}$$

Writing in terms of inner product,

$$\mathbb{E}[\mathbb{E}[w_{n+1}|\mathcal{F}_n]^2] = \|\mathbb{E}[w_{n+1}|\mathcal{F}_n]\|^2 \tag{A.9}$$

The conditional expectation of w_{n+1} given $\mathcal{F}_n$ is equivalent to the orthogonal projection of w_{n+1} onto closed subspace $\mathcal{F}_n$. Let $P : V \to V$ be the projection operator. Then $\mathbb{E}[w_{n+1}|\mathcal{F}_n] = P_{\mathcal{F}_n}(w_{n+1})$. By the property of orthogonal projection,

$$\|w_{n+1}\|^2 = \|P_{\mathcal{F}_n}(w_{n+1})\|^2 + \|w_{n+1} - P_{\mathcal{F}_n}(w_{n+1})\|^2. \tag{A.10}$$

By the projection property (A.10), $\|w_{n+1} - P_{\mathcal{F}_n}(w_{n+1})\|^2 \geq 0$, $\|w_{n+1}\|^2 = \mathbb{E}[(w_{n+1})^2] \geq norm P_{\mathcal{F}_n}(w_{n+1})^2 = \mathbb{E}[\mathbb{E}[w_{n+1}|\mathcal{F}_n]^2]$.

B Bibliography

References

[1] A. Blake et al. "Statistical models of visual shape and motion". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 356.1740 (1998), pp. 1283–1302.

[2] N. Chopin et al. "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference". In: *The Annals of Statistics* 32.6 (2004), pp. 2385–2411.

[3] E. Çınlar. *Probability and stochastics*. Vol. 261. Springer Science & Business Media, 2011.

[4] A. Doucet and A. M. Johansen. "A tutorial on particle filtering and smoothing: Fifteen years later". In: *Handbook of nonlinear filtering* 12.656-704 (2009), p. 3.

[5] B. Fristedt and L. Gray. *A Modern Approach to Probability Theory*. 1996.

[6] J. Geweke. "Bayesian inference in econometric models using Monte Carlo integration". In: *Econometrica: Journal of the Econometric Society* (1989), pp. 1317–1339.

[7] N. J. Gordon, D. J. Salmond, and A. F. Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE proceedings F (radar and signal processing)*. Vol. 140. 2. IET. 1993, pp. 107–113.

[8] H. He, R. Xiong, and J. Fan. "Evaluation of lithium-ion battery equivalent circuit models for state of charge estimation by an experimental approach". In: *energies* 4.4 (2011), pp. 582–598.

[9] A. Kong, J. S. Liu, and W. H. Wong. "Sequential imputations and Bayesian missing data problems". In: *Journal of the American statistical association* 89.425 (1994), pp. 278–288.

[10] J. S. Liu and R. Chen. "Blind deconvolution via sequential imputations". In: *Journal of the american statistical association* 90.430 (1995), pp. 567–576.

[11] J. S. Liu and R. Chen. "Sequential Monte Carlo methods for dynamic systems". In: *Journal of the American statistical association* 93.443 (1998), pp. 1032–1044.

[12] Q. Miao et al. "Remaining useful life prediction of lithium-ion battery with unscented particle filter technique". In: *Microelectronics Reliability* 53.6 (2013), pp. 805–810.

[13] J. Neto. *Statistical Computation and Simulation*. http://www.di.fc.ul.pt/~jpn/r/ECS/index.html#use-of-importance-sampling-to-compute-an-integral. May 2014.

[14] R. D. Pang. *Advanced Statistical Computing*. https://bookdown.org/rdpeng/advstatcomp/importance-sampling.html. July 2018.

[15] A. Papanicolaou. *Taylor Approximation and the Delta Method*. http://www.stat.rice.edu/~dobelman/notes_papers/math/TaylorAppDeltaMethod.pdf. Apr. 2009.

[16] A. Smith. *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[17] A. Tulsyan et al. "State-of-charge estimation in lithium-ion batteries: A particle filter approach". In: *Journal of Power Sources* 331 (2016), pp. 208–223.

[18] D. W. Zimmerman. "Probability spaces, Hilbert spaces, and the axioms of test theory". In: *Psychometrika* 40.3 (1975), pp. 395–412.