SMC-menetelmät

Laskennallinen tilastotiede - Harjoitustyö

Lasse Rintakumpu

10.5.2021

Sisällys

1 Johdanto **2**
1.1 Suodinongelma . . . 2
1.2 Historiaa . . . 3

2 Monte Carlo -menetelmistä **4**
2.1 Monte Carlo -approksimaatio . . . 4
2.2 Tärkeytysotanta . . . 5

3 Bayesilainen suodin **6**

4 SIR-algoritmi **6**
4.1 Parametrien valinta . . . 8
4.1.1 Otoskoon N valinta . . . 8
4.1.2 Uudelleenotantamenetelmän valinta . . . 8
4.1.3 Ehdotusjakauman valinta . . . 9
4.2 Konvergenssituloksia . . . 10
4.3 Marginaalijakauma . . . 10
4.4 Aikakompleksisuus . . . 10

5 Lopuksi **11**

1 Johdanto

SMC-menetelmät (sequential Monte Carlo -mentelmät) ovat joukko Monte Carlo -algoritmeja, joiden avulla voidaan ratkaista ns. suodinongelma, kun ongelma on epälineaarinen ja/tai ongelmaan liittyvä kohina ei noudata normaalijakaumaa. SMC-menetelmille on lukuisia sovellutuksia esimerkiksi Bayesilaisessa tilastotieteessä, fysiikassa ja robotiikassa.

Tämän tutkielman tavoitteena on esittää pääpiirteittäin SMC-menetelmien teoria sekä joitakin menetelmäperheeseen kuuluvia algoritmeja. Tutkielman ensimmäisessä alaluvussa kuvataan yleisellä tasolla sekä suodinongelma että sen ratkaisujen historiaa. Toisessa alaluvussa käsitellään joitakin Monte Carlo -menetelmiin liittyviä yleisiä tuloksia. Kolmannessa alaluvussa esitellään Bayesilainen viitekehys suodinongelmalle, jonka pohjalta neljännessä alaluvussa kuvataan SIR-algoritmina tunnetu SMC-menetelmä.

Tutkielman esitykset seuraavat erityisesti Simo Särkän kirjaa *Bayesian Filtering and Smoothing* (2013), Fredrik Gustafssonin artikkelia "Particle Filter Theory and Practice with Positioning Applications" (2010) sekä Olivier Cappén, Simon J. Godsillin ja Eric Moulines'n artikkelia "An overview of existing methods and recent advances in sequential Monte Carlo" (2007).

1.1 Suodinongelma

Stokastisten prosessien teoriassa suodinongelmaksi kutsutaan tilannetta, jossa halutaan muodostaa keskineliövirheen mielessä paras mahdollinen estimaatti jonkin järjestelmän tilan arvoille, kun ainoastaan osa tiloista voidaan havaita ja/tai havaintoihin liittyy kohinaa. Tavoitteena on toisin sanoen laskea jonkin prosessin posteriorijakauma kyseisten havaintojen perusteella. Ongelmaa havainnollistaa kaavio (1).

$$\begin{array}{ccccccc} x_1 & \longrightarrow & x_2 & \longrightarrow & x_3 & \longrightarrow & \dots \quad \text{piilossa olevat tilat} \\ \downarrow & & \downarrow & & \downarrow & & \\ y_1 & & y_2 & & y_3 & \dots & \text{havainnot} \end{array} \tag{1}$$

Tässä tutkielmassa keskitytään erityisesti epälineaarisen ns. Markovin piilomallin posteriorijakauman Bayesilaiseen ratkaisuun. Ongelmassa tiedetään, miten havaitut muuttujat y_k kytkeytyvät "piilossa oleviin" muuttujiin x_k sekä osataan sanoa jotain tilamuuttujien todennäköisyyksistä. Oletetaan myös, että piilossa oleville tiloille X_k pätee Markov-ominaisuus, jolloin kutakin hetkeä seuraava tila x_{k+1} riippuu menneistä tiloista $x_{1:k}$ ainoastaan tilan x_k välityksellä. Lisäksi havaittu tila y_k riippuu tiloista x_k ainoastaan jonkin x_k:n funktion kautta. Kun aika-avaruus on diskreetti ja merkitään ajanhetkellä $k = 1, \dots, t$ piilossa olevan prosessin tilaa x_k ja havaittua prosessia y_k, saadaan mallit

$$x_{k+1} = f(x_k, \nu_k) \tag{2}$$
$$y_k = h(x_k) + e_k. \tag{3}$$

Lisäksi tiedetään prosessin alkuhetken jakauma $x_1 \sim p_{x_1}$, tähän liittyvän kohinaprosessin jakauma $\nu_k \sim p_{\nu_k}$ sekä malliin y_k liittyvä kohina $e_k \sim p_{e_k}$. Koska SMC-algoritmit pyrkivät ratkaisemaan juurikin epälineaarisen, ei-Gaussisen suodinongelman, voivat funktiot $f(\cdot)$ ja $h(\cdot)$ olla tässä epälineaarisia eikä kohinan tarvitse olla normaalijakautunutta.

Mallit voidaan esittää myös yleisemmässä jakaumamuodossa

$$x_{k+1} \sim p(x_{k+1}|x_k), \tag{4}$$
$$y_k \sim p(y_k|x_k). \tag{5}$$

Tutkielman teoriaosassa käytetään ensisijaisesti yhtälöiden (4) ja (5) muotoilua. Empiirisessä osassa palataan yhtälöiden (2) ja (3) muotoiluun.

Suodinongelmaa lähellä on myös ns. tasoitusongelma (smoothing problem), jossa ollaan kiinnostuneita prosessin x_k posteriorijakaumasta $p(x_k|y_k)$ jokaisena ajanhetkenä $1, \dots, k$ ei ainoastaan haluttuna ajanhetkenä

k. Tämä tutkielma keskittyy yksin suodinongelman ratkaisemiseen, mutta huomioitavaa on, että SMC-algoritmit näyttävät ratkaisevan tasoitusongelman ilmaiseksi. Tähän liittyy kuitenkin joidenkin mallien kohdalla mahdollista epätarkkuutta, joten tarvittaessa tasoitusongelma pitää ratkaista erikseen.

1.2 Historiaa

Tämä alaluku esittää pääpiirteittään suodinongelmalle esitettyjen ratkaisujen historian. Lineaarisen suodinongelman osalta alaluku noudattaa Dan Crisanin artikkelia "The stochastic filtering problem: a brief historical account" (2014) sekä Mohinder S. Grewalin ja Angus P. Andrewsin artikkelia "Applications of Kalman Filtering in Aerospace 1960 to the Present" (2010). SMC-menetelmien osalta lähteenä toimii Cappé &al (2007).

Suodinongelma nousi esille insinööritieteiden sekä sotateollisuuden käytännön ongelmista 2. maailmansodan aikana, vaikkakin suodinongelman diskreetin ajan ratkaisut juontavat jo Andrei N. Kolmogorovin 30-luvun artikkeleihin. Jatkuvan ajan tilanteessa ensimmäisen optimaalisen, kohinan sallivan suotiman esitti matemaatikko, kybernetiikan kehittäjä Norbert Wiener. Wiener-suotimena tunnettua ratkaisuaan varten Wiener muotoili seuraavat kolme ominaisuutta, jotka prosessin X estimaatin $\hat{X}_t$ pitää toteuttaa.

1. *Kausaliteetti*: X_t tulee estimoida käyttäen arvoja Y_s, missä $s \leq t$.
2. *Optimaalisuus*: X_t:n estimaatin $\hat{X}_t$ tulee minimoida keskineliövirhe $\mathbb{E}[(X - \hat{X}_t)^2]$.
3. *On-line -estimointi*: Estimaatin $\hat{X}_t$ tulee olla saatavissa minä hyvänsä ajanhetkenä t.

Wiener sovelsi ratkaisussaan stationaaristen prosessien spektriteoriaa. Tulokset julkaistiin salaisina Yhdysvaltojen asevoimien tutkimuksesta vastanneen National Defense Research Committeen (NDRC) raportissa vuonna 1942. Tutkimus tunnettiin sodan aikana lempinimellä "Keltainen vaara" sekä painopaperinsa värin että vaikeaselkoisuutensa vuoksi. Myöhemmin Wiener esitti tuloksensa julkisesti kirjassaan *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (1949). Wienerin alkuperäiset kolme perusperiaatetta päteveät edelleen kaikille suodinongelman ratkaisuille, myös SMC-menetelmille.

Kenties tärkein ja varmasti tunnetuin lineaariseen suodinongelman ratkaisu on Kalman-suodin. Suotimen kehittivät R.E. Kalman ja R.S. Bucy 1950- ja 60-lukujen taitteessa Yhdysvaltain kylmän sodan kilpavarustelutarpeisiin perustetussa Research Institute for Advanced Studies -tutkimuslaitoksessa (RIAS). Kalman-suodin on suodinongelman diskreetin ajan ratkaisu, kun taas Kalman-Bucy-suodin on jatkuvan ajan ratkaisu. Kohinan ollessa normaalijakautunutta on Kalman-suodin Wiener-suotimen tavoin lineaarisen suodinongelman optimaalinen ratkaisu. Wiener-suotimella ja Kalman-suotimella on kuitenkin erilaiset oletukset, minkä vuoksi erityisesti säätö- ja paikannussovelluksissa Kalman-suotimen käyttö on luontevampaa. Suotimien oletuksia ja oletusten välisiä eroja ei käsitellä tässä tutkielmassa, kuten ei käsitellä myöskään Kalman-suotimen formaalia yhteyttä SMC-menetelmiin.

Kalman-suodinta voidaan soveltaa myös epälineaarisessa tapauksessa, kunhan suodinongelman funktiot $f(\cdot)$ ja $h(\cdot)$ ovat derivoituvia ja niihin liittyvä kohina oletetaan normaalijakautuneeksi. Tätä rataisua kutsutaan laajennetuksi Kalman-suotimeksi (extended Kalman filter, EKF). Suodin kehitettiin 60-luvulla NASA:n Apollo-ohjelman tarpeisiin, vaikkakin itse avaruusalusten laitteistot hyödynsivät lentoratojen laskennassa Kalman-suotimen perusversiota. Laajennetun Kalman-suotimen toimintaperiaate perustuu epälineaaristen funktioiden linearisointiin Taylorin kehitelmän avulla kulloisenkin estimaatin ympärillä. Laajennettu Kalman-suodin on erityisesti paikannussovellusten *de facto* -suodinstandardi, mutta suodin ei kuitenkaan ole epälineaarisen ongelman optimaalinen estimaattori.

Kalman-suotimesta on lisäksi olemassa lukuisia muita epälineaarisiin ongelmiin soveltuvia laajennuksia, muun muassa paikkaratkaisun Kalman-suodin (position Kalman filter, PKF), hajustamaton Kalman-suodin (unscented Kalman filter, UKF) ja tilastollisesti linearisoitu Kalman-suodin (statistically linearized Kalman filter, SLF). Kuitenkin jos prosessin X mallia ei tunneta tarkasti tai kohinaa ei voida olettaa normaalijakautuneeksi, ovat sekventiaaliset Monte Carlo -menetelmät Kalman-suotimen johdannaisia parempia ratkaisuja. Vaikka tila-avaruuden dimensioiden kasvaessa kasvaa myös SMC-menetelmien vaatima laskentateho, ovat SMC-menetelmät aina sitä parempia mitä epälineaarisempia mallit ovat ja mitä kauempana normaalijakaumasta kohina on. Viimeisten vuosikymmenten aikana myös laskennan teho on kasvanut merkittävästi samalla

kun laskennan hinta on vastaavasti romahtanut, mikä puoltaa Monte Carlo -menetelmien käyttöä entistä useammissa ongelmissa.

Joitakin suodinongelman rekursiivisia Monte Carlo -ratkaisuja löytyy jo 1950–70-luvuilta, erityisesti säätöteoriaan piiristä. Olennainen nykyalgoritmeihin periytynyt oivallus varhaisissa suodinalgoritmeissa oli tärkeytysotannan käyttö halutun jakaumaestimaatin laskennassa. Tärkeytysotanta-algoritmiin voidaan turvautua, kun emme pysty suoraan tekemään havaintoja jostakin jakaumasta p ja teemme sen sijaan havaintoja jakaumasta q, joita painotamme niin, että tuloksena saadaan jakauman p harhaton estimaatti. Algoritmi on kuvattu tarkemmin tutkielman alaluvussa 2.

Tärkeysotantaa käyttävä suodinongelman ratkaiseva SIS-algoritmi (sequential importance sampling) ei kuitenkaan vielä 70-luvulla löytänyt suurta käytännön suosiota. Osin tämä johtui puutteellisesta laskentatehosta, mutta algoritmi kärsi myös otosten ehtymisenä (sample impoverishment) tunnetusta ongelmasta. Monissa ongelmissa SIS-algoritmia käytettäessä suuri osa painoista päätyy vain tietyille partikkeleille, jolloin vastaavasti suuri osa partikkeleista ei enää estimoi haluttua jakaumaa. Tähän ongelmaan palataan myöhemmin.

Merkittävän ratkaisun ehtymisongelmaan esittivät Gordon, Salmond ja Smith artikkelissaan "Novel approach to nonlinear/non-Gaussian Bayesian state estimation" (1993). Artikkelin ratkaisu kulki nimellä "bootstrap filter", saapasremmisuodin. Saapasremmisuodin vältti ehtymisen uudellenotannalla, jossa matalapainoiset partikkelit korvattiin otoksilla korkeapainoisemmista partikkeleista. Ratkaisussa painot eivät myöskään riipu partikkelien aiemmista poluista vaan ainoastaan havaintojen uskottavuusfunktiosta. Vastaavaa ratkaisua käytetään tämän tutkielman uudemmassa SIR-algoritmissa (sampling importance resampling), jossa myös uudelleenotantaan sovelletaan tärkeytysotantaa.

SMC-menetelmissä stokastisen prosessin posteriorijakauman esittämiseen käytettyjä otoksia kutsutaan myös partikkeleiksi ja menetelmiä hiukassuotimiksi. Erityisesti myöhemmin esitettävää SIR-algoritmia kutsutaan usein hiukkassuotimeksi. Tässä tutkielmassa pyritään korostamaan suotimien yhteyttä Monte Carlo -algoritmeihin ja käytetään siksi yleisempää termiä SMC-menetelmät. Termiä hiukkassuodin käytti ensimmäisen kerran Del Moral artikkelissa "Nonlinear Filtering: Interacting Particle Resolution" (1996). SMC-menetelmät termiä Liu ja Chen artikkelissa "Sequential Monte Carlo Methods for Dynamic Systems" (1998).

2 Monte Carlo -menetelmistä

Tässä alaluvussa kuvataan lyhyesti SMC-menetelmissä käytettävien Monte Carlo -menetelmien perusperiaate todennäköisyysjakauman estimoinnissa. Lisäksi esitetään tärkeytysotanta-algoritmi (importance sampling), jonka tarkoituksena on estimoida harhattomasti jakaumaa $p(x|y_{1:k})$, josta emme voi suoraan tehdä otoksia, mutta jota voimme approksimoida toisella jakaumalla q. Esitykset noudattavat Särkkää (2013).

2.1 Monte Carlo -approksimaatio

Bayesilaisessa päättelyssä ollaan yleisesti kiinnostuttu laskemaan johonkin posteriorijakaumaan p liittyvää odotusarvoa

$$\mathbb{E}[g(x)|y_{1:k}] = \int g(x)p(x|y_{1:k})dx, \tag{6}$$

missä g on tila-avaruuden mielivaltainen funktio ja $p(x|y_{1:t})$ on havaintoihin $y_1, \ldots, y_k$ liittyvä x:n posterioritiheysjakauma. Odotusarvo on laskettavissa suljetussa muodossa vain harvoissa tapauksissa, suodinongelman kohdalla silloin, kun kyseessä on lineaarinen ja Gaussinen malli. Odotusarvoa voidaan kuitenkin approksimoida niin sanoituilla Monte Carlo -menetelmillä. Menetelmien perusperiaate on tehdä riippumattomia otoksia estimoitavasta jakaumasta ja laskea haluttu odotusarvo otosten avulla. Jos teemme N otosta jakaumasta $x^i \sim p(x|y_{1:t})$, missä $i = 1, \ldots, N$ saadaan näiden otosten avulla laskettua odotusarvon estimaatti

$$\mathbb{E}[g(x)|y_{1:k}] \simeq \frac{1}{N}\sum_{i=1}^{N} g(x^i). \tag{7}$$

Monte Carlo -estimaatti konvergoi keskeisen raja-arvolauseen nojalla ja sen estimointivirheen voidaan osoittaa olevan luokkaa $O(\frac{1}{\sqrt{N}})$ riippumatta tilamuuttujan x dimensiosta. SMC-menetelmät hyödyntävät Monte Carlo -estimointia sekventiaalisesti, jolloin estimaatti lasketaan rekursiivisesti kullekin ajanhetkelle $k = 1, \ldots, t$. Tähän palataan alaluvuissa 3 ja 4.

2.2 Tärkeytysotanta

Tilanteessa, jossa Monte Carlo -otoksia ei voida tehdä suoraan jakaumasta p, voidaan hyödyntää jakaumaa p approksimoivaa tärkeytys- tai ehdotusjakaumaa $q(x|y_{1:k})$ sekä ns. tärkeytysotantaa. Oletetaan, että tunnetaan priorijakauma $p(x)$ ja on olemassa havaintomalli $p(y_{1:k}|x)$ sekä valittu ehdotusjakauma $q(x|y_{1:k})$, josta voidaan tehdä otoksia. Ehdotusjakaumalta edellytetään lisäksi, että sen kantaja on suurempi tai yhtä suuri kuin jakauman $p(x|y_{1:k})$ ja että se saa nollasta poikkeavia arvoja kaikkialla missä $p(x|y_{1:k})$ saa nollasta poikkeavia arvoja. Kirjoitetaan halutun posteriorijakauman odotusarvo integraalina

$$\int g(x)p(x|y_{1:k})dx = \int g(x)\frac{p(x|y_{1:k})}{q(x|y_{1:k})}q(x|y_{1:k})dx, \tag{8}$$

jolle voidaan muodostaa Monte Carlo -approksimaatio tekemällä N otosta jakaumasta $x^i \sim q(x|y_{1:k})$. Muodostetaan näin odotusarvo

$$\mathbb{E}[g(x)|y_{1:k}] \simeq \frac{1}{N}\sum_{i=1}^{N}\frac{p(x^i|y_{1:k})}{q(x^i|y_{1:k})}g(x^i) = \sum_{i=1}^{N} w^i g(x^i), \tag{9}$$

missä $g(x)$ on jokin estimoinnissa hyödyllinen, mielivaltainen funktio. Tutkielmassa käytetty notaatio x_k^i viittaa ajanhetken k partikkeliin i, missä $i = 1, \ldots, N$. Tärkeytysotantaa kuvaa nyt algoritmi (1). Kun posteriorijakauman estimaatti muodostetaan kyseisellä algoritmilla voidaan tulos kirjoittaa

$$\hat{p}(x|y_{1:k}) = \sum_{i=1}^{N} w^i \delta(x - x^i), \tag{10}$$

missä $\delta(x)$ on Diracin deltafunktio.

Algoritmi 1: Tärkeytysotanta

begin
  for $i = 1, 2, \ldots, N$ **do**
    begin
      Otetaan N otosta ehdotusjakaumasta $x^i \sim q(x|y_{1:k})$.
    begin
      Lasketaan normalisoimattomat painot $w_*^i = p(y_{1:k}|x^i)p(x^i)/q(x^i|y_{1:k})$.
      ja normalisoidut painot $w^i = w_*^i / \sum_{j=1}^{N} w_*^j$.
    begin
      Estimoidaan p laskemalla tiheydelle approksimaatio $\mathbb{E}[g(x)|y_{1:k}] \simeq \sum_{i=1}^{N} w^i g(x^i)$.

3 Bayesilainen suodin

Suodinongelmassa ollaan kiinnostuttu tilavektorin posteriorijakauman $p(x_k|y_{1:k})$ estimoinnista. Tässä alaluvussa käydään läpi epälineaarisen tilanteen yleinen rekursiivinen, Bayesilainen posteriorijakauman laskenta. Tällaista suodinongelman ratkaisua kutsutaan myös Bayesilaiseksi suotimeksi. Koska epälineaarisessa, ei-normaalijakautuneessa tilanteessa rekursiota ei voida laskea analyyttisesti, pitää estimoinnissa käyttää numeerisia menetelmiä. SMC-menetelmissä tämä tarkoittaa jakauman sekventiaalista Monte Carlo -approksimointia, jonka toteutus esitetään alaluvun 4 algoritmissa. Molemmat esitykset noudattavat Gustafssonia (2010).

Bayesilainen ratkaisu tilavektorin posteriorijakauman estimaatille $\hat{p}(x_k|y_{1:k})$ saadaan seuraavalla rekursiolla (käydään läpi jokaiselle ajanhetkelle $k = 1, \ldots, t$). Lasketaan ensin

$$p(x_k|y_{1:k}) = \frac{p(y_k|x_k)p(x_k|y_{1:k-1})}{p(y_k|y_{1:k-1})}, \tag{11}$$

joka saadaan suoraan Bayesin kaavasta $P(A|B) = P(B|A)P(A)/P(B)$. Normalisointivakio lasketaan integraalina

$$p(y_k|y_{1:k-1}) = \int_{\mathbb{R}^{n_x}} p(y_k|x_k)p(x_k|y_{1:k-1})\, dx_k, \tag{12}$$

joka saadaan kokonaistodennäköisyyskaavasta $P(A) = \mathbb{E}[P(A|X)] = \int_{-\infty}^{\infty} P(A|X = x)f_X(x)\, dx$. Merkintä R^{n_x} vastaa tässä piilossa olevan tilavektorin dimensiota n.

Lopuksi lasketaan päivitysaskel ajalle, joka saadaan edelleen kokonaistodennäköisyydellä

$$p(x_{k+1}|y_{1:k}) = \int_{\mathbb{R}^{n_x}} p(x_{k+1}|x_k)p(x_k|y_{1:k})\, dx_k\,. \tag{13}$$

Rekursion avulla voimme laskea jakauman $p(x_k|y_{1:k})$ estimaatti käymällä rekursion läpi k kertaa.

4 SIR-algoritmi

Tässä alaluvussa esitetään SMC-menetelmiin kuuluva SIR-algoritmi Bayesilaisen, epälineaarisen suodinongelman ratkaisemiseksi. Algoritmi on numeerinen toteutus alaluvussa 3 kuvatusta Bayesilaisesta suotimesta. Esitetty algoritmi perustuu Gustafssoniin (2010).

Algoritmi alustetaan jakaumasta $x_1^i \sim p_{x_0}$ generoiduilla N-kappaleella partikkeleita. Jokaiselle partikkelille annetaan alustuksessa sama paino $w_{1|0}^i = 1/N$. Algoritmi suoritetaan jokaiselle partikkelille $i = 1, 2, \ldots, N$ jokaisella ajanhetkellä $k = 1, 2, \ldots, t$.

Seuraava toistetaan jokaiselle ajanhetkelle $k = 1, 2, \ldots, t$. Algoritmin ensimmäisessä vaiheessa päivitetään painot yhtälön 14 mukaan.

$$w_{k|k}^i = \frac{1}{c_k} w_{k|k-1}^i p(y_k|x_k^i). \tag{14}$$

Tämä vastaa yllä esitetyn Bayes-suotimen päivitysvaihetta (11). Normalisointipaino c_k lasketaan puolestaan yhtälöstä (15), mikä vastaa Bayes-suotimen normalisointivakion laskemista (12) ja asettaa painojen summaksi $\sum_{i=1}^N w_{k|k}^i = 1$.

$$c_k = \sum_{i=1}^{N} w_{k|k-1}^i p(y_k|x_k^i). \tag{15}$$

Seuraavassa vaiheessa estimoidaan p laskemalla tiheyden $p(x_{1:k}|y_{1:k})$ Monte Carlo -estimaatti yhtälön (16) perusteella

$$\hat{p}(x_{1:k}|y_{1:k}) = \sum_{i=1}^{N} w_{k|k}^{i} \delta(x_{1:k} - x_{1:k}^{i}). \tag{16}$$

Tämän jälkeen suoritetaan valinnainen uudelleenotanta. Uudelleenotanta voidaan tehdä jokaisella askeleella tai efektiivisen otoskoon perusteella alla kuvatun kynnysarvoehdon $\hat{N}_{eff} < N_{th}$ täyttessä, jolloin uudelleenotantaa kutsutaan adaptiiviseksi uudelleenotannaksi. Tällaista uudelleenotantaa hyödynnetään esitetyssä algoritmissa (2) . Uudelleenotantaa tarkastellaan lähemmin alaluvussa 4.1.2. Lopuksi päivitetään aika (jos $k < t$) ja luodaan uudet ennusteet partikkeleille ehdotusjakaumasta (17)

$$x_{k+1}^{i} \sim q(x_{k+1}|x_{k}^{i}, y_{k+1}) \tag{17}$$

ja päivitetään partikkelien painot tärkeytysotannalla (18), sen mukaan kuinka todennäköisiä partikkelien ennusteet ovat

$$w_{k+1|k}^{i} = w_{k|k}^{i} \frac{p(x_{k+1}^{i}|x_{k}^{i})}{q(x_{k+1}^{i}|x_{k}^{i}, y_{k+1})}. \tag{18}$$

Vaiheet 17 ja 18 vastaavat Bayes-suotimen aikapäivitystä (13).

Alla käsitellään algoritmiin liittyvän uudelleenotantamenetelmän, partikkelien määrän ja ehdotusjakauman valinta. Lopuksi esiteetään algoritmin konvergenssia, marginaalijakaumaa sekä aikakompleksisuutta koskevia tuloksia.

Algoritmi 2: SIR

Result: Posteriorijakauman $p(x_{1:k}|y_{1:k})$ estimaatti.
Data: Havainnot y_k. Generoitu $x_1^i \sim p_{x_0}$ missä $i = 1, \ldots, N$ ja jokainen partikkeli saa saman painon $w_{1|0}^i = 1/N$.

```
begin
    for k = 1, 2, ..., t do
        for i = 1, 2, ..., N do
            begin
                Päivitetään painot w_{k|k}.
            begin
                Estimoidaan p laskemalla tiheydelle approksimaatio p̂(x_{1:k}|y_{1:k}) = Σ_{i=1}^N w_{k|k}^i δ(x_{1:k} − x_{1:k}^i).
        begin
            Lasketaan efektiivinen otoskoko N̂_eff.
        if N̂_eff < N_th then
            begin
                Otetaan uudet N otosta palauttaen joukosta {x_{1:k}^i}_{i=1}^N, missä otoksen i todennäköisyys on
                w_{k|k}^i.
        if k < t then
            begin
                Aikapäivitys.
                Luodaan ennusteet partikkeleille ehdotusjakaumasta x_{k+1}^i ~ q(x_{k+1}|x_k^i, y_{k+1}),
                päivitetään partikkelien painot tärkeytysotannalla.
```

4.1 Parametrien valinta

Ennen algoritmin suorittamista valitaan ehdotusjakauma $q(x_{k+1}|x_{1:k}, y_{k+1})$, uudelleenotantamenetelmä sekä partikkelien määrä N. Ehdotusjakauman ja uudelleenotantamenetelmän valinnassa tärkeimpänä päämääränä on välttää otosten ehtymistä, kun taas partikkelien määrä säätelee kompromissia algoritmin suorituskyvyn ja tarkkuuden välillä.

4.1.1 Otoskoon N valinta

Yleispätevää sääntöä otoskoon/partikkelien lukumäärän N valinnalle on vaikeaa antaa, sillä vaadittava estimointitarkkuus riippuu usein käsillä olevasta ongelmasta. Gordon &al. (1993) esittävät kuitenkin kolme tekijää, jotka vaikuttavat partikkelien lukumäärän valintaan

a. tila-avaruuden ulottuvuuksien lukumäärä n_x,
b. tyypillinen päällekäisyys priorin ja uskottavuuden välillä
c. sekä tarvittava aika-askeleiden lukumäärä.

Ensimmäisen tekijän vaikutus on selvä. Mitä useammassa ulottuvuudessa otantaa tarvitsee tehdä, sen korkeammaksi on N asetettava, jotta jokainen ulottuvuus pystytään kattamaan. Tekijät (b) ja (c) puolestaan seuraavat uudelleenotannasta. Jos se osa tila-avaruutta, jossa uskottavuus $p(y_k|x_k)$ saa merkittäviä arvoja on pieni verrattuna siihen osaan, jossa priorijakauma $p(x_k|y_{1:k-1})$ saa merkittäviä arvoja, suuri osa partikkeleista saa pieniä painoja eikä näin valikoidu uudelleenotantaan.

Yleisesti ottaen N kannattaa asettaa sellaiseksi, että se paitsi tuottaa riittävän tarkan estimaatin, on se käytettävissä olevan laskentatehon sekä vaadittavan laskentanopeuden kannalta järkevää.

4.1.2 Uudelleenotantamenetelmän valinta

Ilman uudelleenotantaa on mahdollista, että algoritmi alkaa kärsiä SIS-algoritmille ominaisesta otosten ehtymisestä. Toisin sanoen kaikki painot alkavat keskittyä vain muutamalle partikkelille eikä algoritmi enää approksimoi tehokkaasti haluttua jakaumaa. Uudelleenotanta tarjoaa osittaisen ratkaisun tähän ongelmaan, mutta hävittää samalla informaatiota ja siten lisää satunnaisotantaan liittyvää epävarmuutta. Yleisesti ottaen uudelleenotanta kannattaa aloittaa vasta siinä vaiheessa algoritmin suorittamista, kun siitä on otosten ehtymisen kannalta hyötyä, esimerkiksi efektiivisen otoskoon laskettua jonkin kynnysarvon alapuolelle (adaptiivinen uudelleenotanta). Efektiivinen otoskoko saadaan laskettua variaatiokertoimesta c_ν kaavalla

$$N_{eff} = \frac{N}{1 + c_\nu^2(w_{k|k}^i)} = \frac{N}{1 + \frac{\mathrm{Var}(w_{k|k}^i)}{(\mathbb{E}[w_{k|k}^i])^2}} = \frac{N}{1 + N^2\mathrm{Var}(w_{k|k}^i)}. \tag{19}$$

Näin laskettu efektiivinen otoskoko maksimoituu ($N_{eff} = N$), kun kaikille painoille pätee $w_{k|k}^i = 1/N$ ja minimoituu ($N_{eff} = 1$), kun $w_{k|k}^i = 1$ todennäköisyydellä $1/N$ ja $w_{k|k}^i = 0$ todennäköisyydellä $(N-1)/N$. Normalisoitujen painojen avulla saadaan effektiiviselle otoskoolle ajanhetkellä k laskennallinen approksimaatio

$$\hat{N}_{eff} = \frac{1}{\sum_{i=1}^N (w_{k|k}^i)^2}. \tag{20}$$

Sekä määritelmälle (19) että (20) pätee $1 \leq \hat{N}_{eff} \leq N$. Yläraja saavutetaan, kun jokaisen partikkelin paino on sama. Alarajalle puolestaan päädytään, kun kaikki paino keskittyy yksittäiselle partikkelille. Tästä saadaan määriteltyä algoritmille SIR-uudelleenotantaehto $\hat{N}_{eff} < N_{th}$. Gustafsson (2010) esittää uudelleenotannan kynnysarvoksi esimerkiksi $\hat{N}_{th} = 2N/3$.

Uudelleenotanta ei muuta approksimoitavan jakauma p odotusarvoa, mutta se lisää jakauman Monte Carlo -varianssia. On kuitenkin olemassa esimerkiksi osittamiseen perustuvia uudelleenotantamenetelmiä, jotka pyrkivät minimoimaan varianssin lisäyksen. Varianssin pienennysmenetelmät jätetään tämän tutkielman ulkopuolelle.

4.1.3 Ehdotusjakauman valinta

Yksinkertaisin muoto ehdotusjakaumalle on $q(x_{1:k}|y_{1:k})$ eli jokaisella algoritmin suorituskerralla käydään läpi koko aikapolku $1:k$. Tämä ei kuitenkaan ole tarkoituksenmukaista, erityisesti jos kyseessä on reaaliaikainen sovellutus. Kirjoitetaan ehdotusjakauma muodossa

$$q(x_{1:k}|y_{1:k}) = q(x_k|x_{1:k-1}, y_{1:k})q(x_{1:k-1}|y_{1:k}). \tag{21}$$

Jos yhtälöstä (21) poimitaan ehdotusjakaumaksi ainoastaan termi $q(x_k|x_{1:k-1}, y_{1:k})$ voidaan tämä kirjoittaa edelleen Markov-ominaisuuden nojalla muotoon $q(x_k|x_{k-1}, y_k)$. Tämä on suodinongelman kannalta riittävää, koska olemme kiinnostuneita posteriorijakaumasta ja arvosta x ainoastaan ajanhetkellä k (tasoitusongelmassa tarvitsisimme koko polun $x_{1:k}$). Alla tarkastellaan edelleen Gustafssonia (2010) seuraten kahta ehdotusjakauman valintatapaa, prioriotantaa (prior sampling) sekä uskottavuusotantaa (likelihood sampling).

Ennen ehdotusjakauman tarkastelua määritellään mallille signaali-kohinasuhde uskottavuuden maksimin ja priorin maksimin välisenä suhteena

$$\text{SNR} \propto \frac{\max_{x_k} p(y_k|x_k)}{\max_{x_k} p(x_k|x_{k-1})}. \tag{22}$$

Yhdistetään lisäksi ehdotusjakaumia varten yhtälöt (14) ja (15), jolloin saadaan painojen päivitys muotoon (23).

$$w^i_{k|k} \propto w^i_{k-1|k-1} \frac{p(y_k|x^i_k)p(x_k|x^{k-1})}{q(x_k|x^i_{k-1}, y_k)} \tag{23}$$

Kun suhde (22) on matala, on prioriotanta luonnollinen valinta. Tässä käytetään ehdotusjakaumana tilavektorin ehdollista prioria eli

$$q(x_k|x_{1:k-1}, y_k) = p(x_k|x^i_{k-1}). \tag{24}$$

Yhtälön (24) perusteella saadaan edelleen prioriotannan painoiksi (25)

$$w^i_{k|k} = w^i_{k|k-1}p(y_k|x^i_k) = w^i_{k-1|k-1}p(y_k|x^i_k). \tag{25}$$

Kun signaali-kohinasuhde on kohtalainen tai korkea, on parempi käyttää ehdotusjakaumana skaalattua uskottavuusfunktiota (27). Tarkastellaan ensin tekijöihin jakoa (26).

$$p(x_k|x^i_{k-1}, y_k) = p(y_k|x_k)\frac{p(x_k|x^i_{k-1})}{p(y_k|x^i_{k-1})} \tag{26}$$

Kun SNR on korkea ja uskottavuusfunktio on integroituva pätee $p(x_k|x^i_{k-1}, y_k) \propto p(y_k|x_k)$, jolloin voidaan asettaa (27)

$$q(x_k|x^i_{k-1}, y_k) \propto p(y_k|x_k). \tag{27}$$

Yhtälön (27) perusteella saadaan edelleen uskottavuusotannan painoiksi (28).

$$w^i_{k|k} = w^i_{k-1|k-1}p(x^i_k|x^i_{k-1}). \tag{28}$$

4.2 Konvergenssituloksia

Alla esitetään kaksi SIR-algoritmiin liittyvää konvergenssitulosta, se kuinka hyvin esitetyllä algoritmilla arvioitu posterioritiheys $\hat{p}(x_{1:k}|y_{1:k})$ approksimoi todellista tiheysfunktiota $p(x_{1:k}|y_{1:k})$ sekä mikä on approksimaation keskineliövirhe. Tulokset noudattavat Crisanin ja Doucet'n artikkeleita "Convergence of Sequential Monte Carlo Methods" (2000) ja "A Survey of Convergence Results on Particle Filtering Methods for Practitioners" (2002).

Konvergenssitulos 1: Kun $N \to \infty$ algoritmille pätee $\forall k$ tulos (29).

$$\hat{p}(x_{1:k}|y_{1:k}) \xrightarrow{a.s.} p(x_{1:k}|y_{1:k}). \tag{29}$$

Konvergenssitulos 2: Keskineliövirheelle pätee asymptoottinen konvergenssi (30).

$$\mathbb{E}(\hat{g}(x_k) - \mathbb{E}(g(x_k)))^2 \leq \frac{p_k \left\| g(x_k) \right\|}{N}, \tag{30}$$

missä g on mikä hyvänsä piilossa olevan tila-avaruuden rajoitettu Borel-mitallinen funktio ($g \in \mathcal{B}(\mathbb{R}^{n_x})$), $\left\| g(\cdot) \right\|$ kyseisen funktion supremum-normi ja p_k jokin äärellinen vakio, jolle pätee ajanhetkestä k riippumatta $p_k = p < \infty$. Konvergenssituloksia ei tämän tutkielman puitteissa todisteta.

4.3 Marginaalijakauma

Edellä kuvattu algoritmi 1 tuottaa approksimaation koko prosessin posteriorijakaumalle $p(x_{1:k}|y_{1:k})$. Jos halutaan tietää ainoastaan posteriorijakauman $p(x_k|y_{1:k})$ estimaatti, voidaan käyttää yksinkertaisesti viimeisestä tilasta x_k laskettua estimaattia

$$\hat{p}(x_k|y_{1:k}) = \sum_{i=1}^{N} w_{k|k}^i \delta(x_k - x_k^i). \tag{31}$$

Toinen, tarkempi vaihtoehto on käyttää laskennassa tärkeytyspainoa

$$w_{k+1|k}^i = \frac{\sum_{j=1}^{N} w_{k|k}^j p(x_{k+1}^i|x_k^j)}{q(x_{k+1}^i|x_k^i, y_{k+1})} \tag{32}$$

painon (18) sijaan. Tällöin jokaisessa aikapäivitysaskeleessa lasketaan painot kaikkien mahdollisten tila-aika-avaruuspolkujen yli. Samoin kuin uudelleenotanta tämä pienentää painojen varianssia.

4.4 Aikakompleksisuus

Algoritmin perusmuodon aikakompleksisuus on $\mathcal{O}(N)$. Uudelleenotantamenetelmän tai ehdotusjakauman valinta ei suoraan vaikuta aikakompleksisuuteen. Sen sijaan marginalisointi tärkeytyspainolla (32) lisää algoritmin aikakompleksisuutta $\mathcal{O}(N) \to \mathcal{O}(N^2)$, koska jokaisen partikkelin kohdalla painot lasketaan jokaisen tila-aika-avaruuspolun yli. On selvää, että erityisesti isoilla otoskoon N arvoilla ei yllä esitetty marginalisointi enää ole mielekästä.

Tällaisia tilanteita varten algoritmista on olemassa $\mathcal{O}(N\log(N))$ -versioita, jotka perustuvat esimerkiksi N:n kappaleen oppimiseen (N-body learning). Näiden algoritmien käsittely jää tämän tutkielman ulkopuolelle, mutta katsauksen algoritmeista ovat esittäneet esimerkiksi Klaas &al. artikkelissa "Toward Practical N^2 Monte Carlo: the Marginal Particle Filter" (2012).

5 Lopuksi

Tässä tutkielmassa on esitetty pääpiirteittäin SMC-menetelmien teoria Bayesilaisessa tilastotieteellisessä viitekehyksessä. Lisäksi tutkielmassa on käyty läpi uudelleenotantaa efektiivisen otoskoon perusteella hyödyntävä SIR-suodinalgoritmi. Lopuksi tutkielmassa on tarkasteltu SIR-algoritmin parametrien valintaan, suorituskykyyn sekä konvergenssiin liittyviä tuloksia.

Viitteet

[1] Cappé, O., Godsill, S. J., Moulines, E., 2007. An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo. *Proceedings of the IEEE*, Vol 95, No 5, 899–924.

[2] Crisan, D. 2014. The stochastic filtering problem: A brief historical account. *Journal of Applied Probability*, Vol 51(A), 13–22.

[3] Crisan, D., Doucet, A. 2000. *Convergence of Sequential Monte Carlo Methods.*

[4] Crisan, D., Doucet A. 2002. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, Vol 50, No 3, 736–746.

[5] Davidson, P., Collin J., Takala J. 2010, Application of particle filters for indoor positioning using floor plans. *2010 Ubiquitous Positioning Indoor Navigation and Location Based Service*, 1–4.

[6] Del Moral, . 1997. Nonlinear filtering: Interacting particle resolution. *Markov Processes and Related Fields*, Vol 2, No 4, 555–580.

[7] Gordon, N.J., Salmond, D.H., Smith A.F.M. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process*, Vol 140, No 2, 107–113.

[8] Grewal, M.S., Andrews, A.P. 2010. Applications of Kalman Filtering in Aerospace 1960 to the Present [Historical Perspectives]. *IEEE Control Systems Magazine*, Vol 30, No 3, 69–78.

[9] Gustafsson, F. 2010. Particle filter theory and practice with positioning applications. *IEEE Aerospace and Electronic Systems Magazine*, Vol 25, No 7, 53–82.

[10] Liu, J., Chen, R. 1998. Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association*, Vol 93, No 443, 1032–1044.

[11] Munoz, D., Enriquez-Caldera, R., Vargas, C., Bouchereau Lara, F. 2009. *Position Location Techniques and Applications.* Elsevier.

[12] Särkkä, S. 2013. *Bayesian Filtering and Smoothing.* Cambridge University Press.