

Kitekintés

Sztochasztika

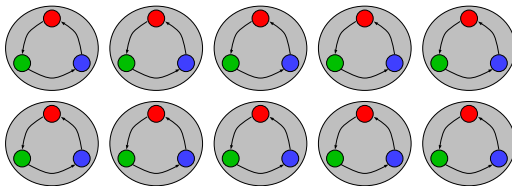
Illés Horváth

2024/12/03

- (1) Markov populációs folyamatok
- (2) Minta generálása adott eloszlásból
- (3) PH (fázis típusú) eloszlások
- (4) Általános Markov-sorra a rendszerben töltött idő eloszlásának kiszámítása Laplace-transzformálttal

Markov populációs folyamatok

Egy folytonos idejű Markov-láncból vegyünk N példányt, amik párhuzamosan zajlanak. Egyetlen Markov-lánc állapottere K méretű, generátora $Q = (r_{ij})_{i,j=1}^K$.



Hogy lehet a rendszer pillanatnyi állapotát leírni?

Egy lehetőség, hogy minden egyes Markov-lánc állapotát nyilvántartjuk („teljes rendszer”). Ekkor az állapottér mérete K^N .

Egy másik lehetőség, hogy csak azt tartjuk számon, hogy az egyes állapotokban mennyien tartózkodnak: jelölje $X_k(t) = X_k^N(t)$ a k állapotban tartózkodó Markov-láncok számát t -kor. Ekkor az állapottér

$$\left\{ (X_1, \dots, X_K) : X_k \in \mathbb{Z}, X_k \geq 0, \sum_{k=1}^K X_k = N \right\},$$

aminek a mérete N^K nagyságrendű. Ez is nagy (bár kisebb, mint a teljes rendszer), viszont jól strukturált.

A legegyszerűbb esetben az egyes Markov-lánckok függetlenül zajlanak. Ilyenkor a Markov-tulajdonság teljesül a teljes rendszerre is, és $X^N(t)$ -re is.

Példa. Legyen $N = 10$ és $K = 3$. $X^N(t)$ -nek az $(5, 3, 2)$ állapotból a lehetséges átmenetei:

- $(5, 3, 2) \rightarrow (4, 4, 2)$, ennek rátája $5r_{12}$, vagy
- $(5, 3, 2) \rightarrow (4, 3, 3)$, ennek rátája $5r_{13}$, vagy
- $(5, 3, 2) \rightarrow (6, 2, 2)$, ennek rátája $3r_{21}$, vagy
- $(5, 3, 2) \rightarrow (5, 2, 3)$, ennek rátája $3r_{23}$, vagy
- $(5, 3, 2) \rightarrow (6, 3, 1)$, ennek rátája $2r_{31}$, vagy
- $(5, 3, 2) \rightarrow (5, 4, 1)$, ennek rátája $2r_{32}$.

Azt szeretnénk, hogy az egyes Markov-láncok ne legyenek függetlenek. Például megengedhetjük, hogy az átmenetráták függjenek a többi Markov-lánc állapotától.

Egy lehetséges választás, ha az r_{ij} ráták az

$$x^N(t) = \frac{X^N(t)}{N}$$

függvényei. $x^N(t)$ az $X^N(t)$ normalizált változata, ami azt mondja meg, hogy a Markov-láncok mekkora része van az egyes állapotokban.

Ezt hívjuk *sűrűségfüggő Markov populációs modellnek*.

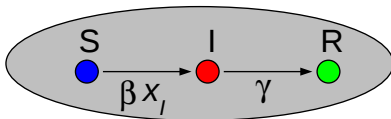
Ilyenkor az egyes Markov-láncok már nem függetlenek. Fontos: a Markov-tulajdonság teljesül $x^N(t)$ -re!

Példa. SIR járványterjedési modell. Egy populációban minden egyed az S, I, R állapotok valamelyikében lehet:

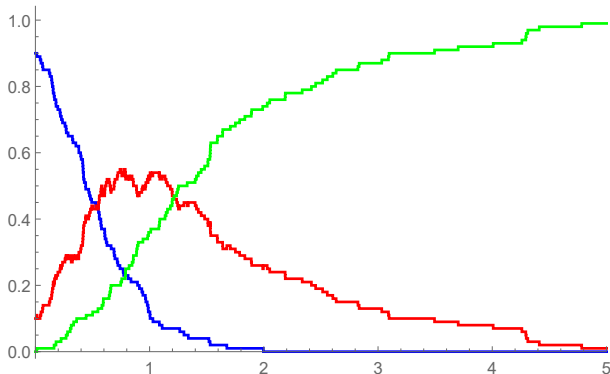
- S (susceptible): egészséges, a fertőzésre fogékony egyed;
- I (infected): fertőzött egyed;
- R (recovered): meggyógyult, a fertőzésre már immunis egyed.

Kétféle átmenet lehetséges:

- $S \rightarrow I$ a megfertőződés, ennek a rátája *arányos a betegek számával*: $r_{S \rightarrow I}(x) = \beta x_I$;
- $I \rightarrow R$ a gyógyulás, rátája konstans: $r_{I \rightarrow R}(x) = \gamma$.

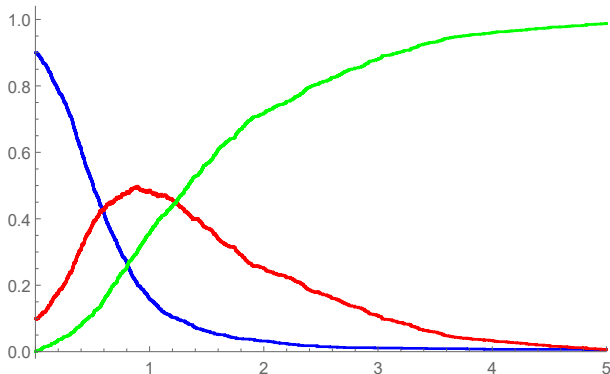


Hogyan néz ki az $x^N(t)$ folyamat egy realizációja?



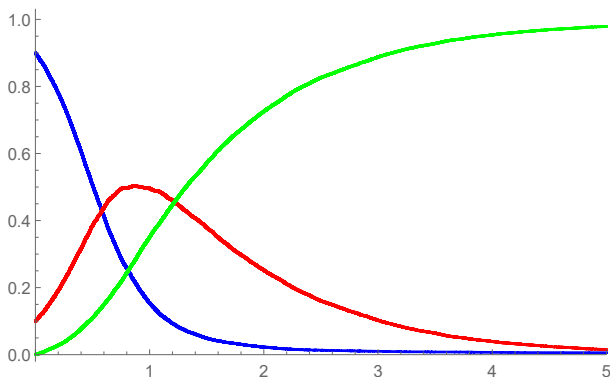
(A paraméterek: $\beta = 5, \gamma = 1, N = 100$, a kezdeti állapot $x(0) = (0.9, 0.1, 0)$.)

Hogyan néz ki az $x^N(t)$ folyamat egy realizációja?



(A paraméterek: $\beta = 5, \gamma = 1, N = 1000$, a kezdeti állapot $x(0) = (0.9, 0.1, 0)$.)

Hogyan néz ki az $x^N(t)$ folyamat egy realizációja?



(A paraméterek: $\beta = 5, \gamma = 1, N = 10000$, a kezdeti állapot $x(0) = (0.9, 0.1, 0)$.)

A fentiek alapján az sejthető, hogy amint $N \rightarrow \infty$, a sűrűségfüggő Markov-folyamatok is koncentrálnak valami determinisztikus folyamatra.

Egy adott sűrűségfüggő Markov populációs folyamathoz tartozó mean-field (térátlag) egyenlet-rendszer a következő:

$$\frac{d}{dt} v_k(t) = \sum_{i \neq k} v_i(t) r_{ik}(v(t)) - \sum_{i \neq k} v_k(t) r_{ki}(v(t)), \quad k = 1, \dots, K$$

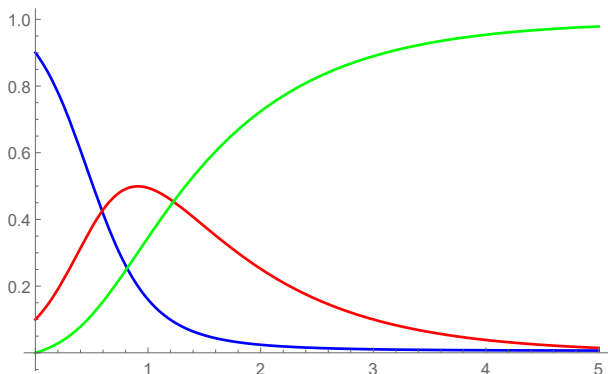
Ez egy K -dimenziós közönséges differenciálegyenlet-rendszer. Ha az $r_{ij}(\cdot)$ függvények folytonosak, akkor a megoldás létezik és egyértelmű.

Példa. Az SIR folyamatra a mean-field egyenlet-rendszer:

$$\begin{aligned}\frac{d}{dt}v_S(t) &= -\beta v_S(t)v_I(t), \\ \frac{d}{dt}v_I(t) &= \beta v_S(t)v_I(t) - \gamma v_I(t), \\ \frac{d}{dt}v_R(t) &= \gamma v_I(t).\end{aligned}$$

(A mean-field magyarul kb. térátlagnak fordítható.)

Hogyan néz ki a mean-field egyenletrendszer megoldása?



(A paraméterek: $\beta = 5, \gamma = 1$, a kezdeti állapot $v(0) = (0.9, 0.1, 0)$.)

Tétel (Kurtz I.)

Tegyük fel, hogy

- *az r_{ij} függvények folytonosan deriválhatóak, és*
- *$x^N(0) \rightarrow v(0)$ valószínűségben, amint $N \rightarrow \infty$.*

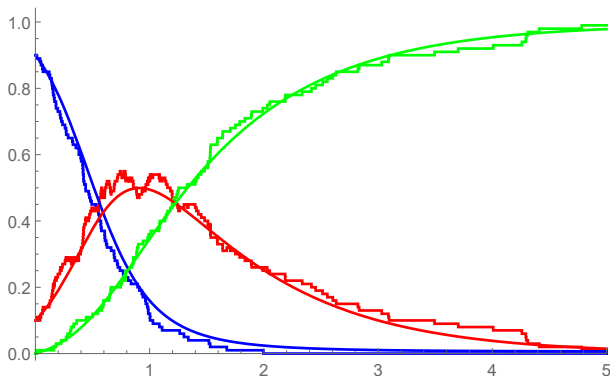
Ekkor a mean-field egyenletrendszer $v(t)$ megoldása egyértelmű, és tetszőleges véges T -re és $\varepsilon > 0$ -ra

$$\lim_{N \rightarrow \infty} \mathbb{P} \left(\max_{0 \leq t \leq T} \max_{1 \leq k \leq K} |v_k(t) - x_k^N(t)| > \varepsilon \right) = 0.$$

Nem biz.

$v(t)$ -t hívjuk a populációs folyamat mean-field limeszének is.

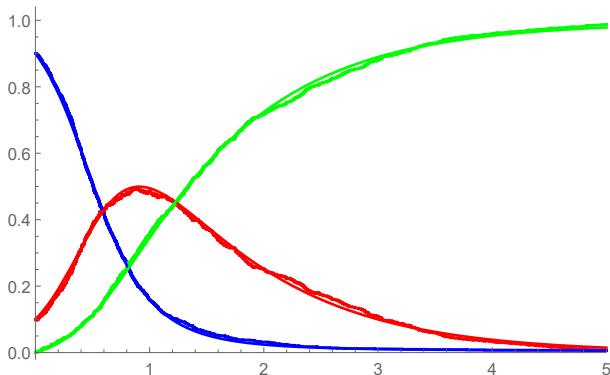
A Kurtz-tétel a gyakorlatban azt jelenti, hogy N növelésével $x^N(t)$ realizációi egyre közelebb kerülnek a mean-field egyenletrendszer megoldásához:



$N = 100$

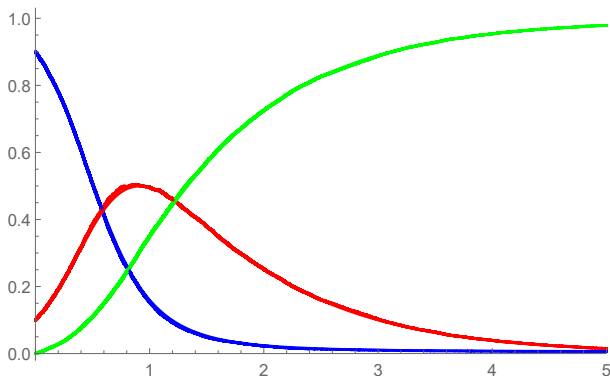
Kurtz-tétel

A Kurtz-tétel a gyakorlatban azt jelenti, hogy N növelésével $x^N(t)$ realizációi egyre közelebb kerülnek a mean-field egyenletrendszer megoldásához:



$N = 1000$

A Kurtz-tétel a gyakorlatban azt jelenti, hogy N növelésével $x^N(t)$ realizációi egyre közelebb kerülnek a mean-field egyenletrendszer megoldásához:



$N = 10000$

Megjegyzések.

A Kurtz-tétel lényegében a Nagy Számok Törvényének általánosítása sűrűségfüggő Markov populációs folyamatokra.

A Centrális Határeloszlás Tétel is teljesül: $x(t)$ fluktuációi $v(t)$ körül $\frac{1}{\sqrt{N}}$ nagyságrendűek, és a fluktuációk eloszlása normális eloszláshoz tart, amint $N \rightarrow \infty$.

Az, hogy $v(t)$ egy közönséges (és nem pl. egy késleltetett) diffegyenlet-rendszer megoldása, annak felel meg, hogy a mean-field limesz memóriamentes, ami éppen azon múlik, hogy az eredeti populációs folyamatra teljesül a Markov-tulajdonság.

A sűrűségfüggő feltevés annak felel meg, hogy minden egyes egyed a teljes populációban tapasztalható sűrűséget érzékeli. Ha van a populációnak belső struktúrája (pl. ismeretségi körök), akkor általában nem ez a helyzet.

További megjegyzések.

Az, hogy $x(t)$ és $v(t)$ is az arányokat írja le a teljes populáción belül, impliciten azt jelenti, hogy minden egyes állapotban a teljes populációval összemérhető számú egyed van.

Például az SIR modellre ez azt jelenti, hogy a mean-field limesz abban a tartományban alkalmazható, amikor a fertőzöttek száma már összemérhető a teljes populációval.

A járvány kezdeti szakaszában, amikor a teljes populációhoz képest sokkal kevesebb fertőzött van, más modellek jobbak (pl. elágazó folyamat).

További megjegyzések.

A sűrűségfüggő feltevés annak felel meg, hogy minden mini Markov-lánc x^N -et, vagyis a teljes globális állapotot érzékeli. Ez a valóságban sok esetben nem így van, hanem az egyes egyedek között van valamiféle struktúra (pl. emberi populációban az ismeretségek), és mindenki csak a saját környezetét érzékeli.

Ilyenkor a rendszer nem sűrűségfüggő Markov-lánc, sőt $x^N(t)$ -re általában nem is teljesül a Markov-tulajdonság. A mean-field limesz (ha egyáltalán létezik), eltérhet Kurtz-tól. Az irodalomban kevés eredmény ismert ilyen esetekre.

Ezzel együtt közelítésnek még ilyen esetben is lehet használni a Kurtz-tételt.

Tétel (Kurtz II.)

Tegyük fel, hogy:

- *az r_{ij} függvények folytonosan deriválhatóak, és*
- *minden N -re $x^N(t)$ -nek egyetlen stacionárius eloszlása van, π_N , és*
- *a $v(t)$ -t definiáló diffegyenlet-rendszernek egyetlen $(v_1, \dots, v_K)$ fixpontja van, és az aszimptotikusan stabil.*

Ekkor a π_N eloszlások koncentrálódnak $(v_1, \dots, v_K)$ -ra, amint $N \rightarrow \infty$.

Nem biz.

Példa. Mi történik, amikor rákattintunk egy YouTube-videóra?

Beküldünk egy igényt, amit majd kiszolgál egy szerver. Na de a YouTube-nak sok szervere van, melyikhez kerül a mi igényünk?

A beérkező igényeket egy külön szerver (“diszpécser”) továbbítja a tényleges kiszolgáló szerverek valamelyikéhez. A cél az igények minél hatékonyabb kiszolgálása, ehhez sok szempontot igénybe vehet, pl.

- a szerverek leterheltsége
- az igény típusa
- fizikai távolság
- stb.

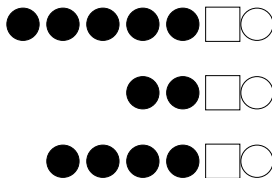
Példa. Mi történik, amikor rákattintunk egy YouTube-videóra?

Beküldünk egy igényt, amit majd kiszolgál egy szerver. Na de a YouTube-nak sok szervere van, melyikhez kerül a mi igényünk?

A beérkező igényeket egy külön szerver (“diszpécser”) továbbítja a tényleges kiszolgáló szerverek valamelyikéhez. A cél az igények minél hatékonyabb kiszolgálása, ehhez sok szempontot igénybe vehet, pl.

- **a szerverek leterheltsége**
- az igény típusa
- fizikai távolság
- stb.

Példa. Hasonlóan működnek a kasszák is egy áruházban. Melyik sorba állnánk be a 3 közül?



Nyilván a legrövidebbhez. Ez a JSQ (Join-Shortest-Queue) néven ismert terhelés-elosztási elv.

Példa. Amikor autós navigációs alkalmazások alternatív útvonalakat keresnek, az is terhelés-elosztás.

Erre a helyzetre fogunk matematikai modellt adni.

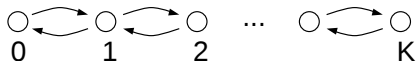
Van egy kiszolgáló rendszerünk, amely N szerverből és egy diszpécserből áll. Az igények mind a diszpécserhez futnak be, amely minden egyes igényt azonnal továbbküld valamelyik szerverhez valamilyen terhelés-elosztási elv alapján (pl. JSQ).

Minden egyes szervernek van egy sora, amiben a beérkező igényeket tárolja. Az egyszerűség kedvéért tegyük fel, hogy a szerverek egyformák és véges K nagyságú bufferrel rendelkeznek. (Egyébként ha többféle szerver van, az sem gond.)

Feltesszük, hogy a bejövő igények Poisson-folyamat szerint érkeznek λN rátával, valamint a kiszolgálás is exponenciális. Ekkor a teljes rendszerre teljesül a Markov-tulajdonság.

Egy szerver állapota az, hogy hány igény van a sorában; az egy szerverhez tartozó Markov-lánc állapottere $\{0, 1, \dots, K\}$.

Minden egyes szerverhez tartozó Markov-lánc egy Markov-sor:



$X_0(t)$ azt jelöli, hogy hány üres szerver van, $X_1(t)$ azt, hogy hány szerverben van 1 igény, és így tovább, egészen $X_K(t)$ -ig. Az

$$x(t) = \frac{X(t)}{N}$$

vektor pedig a megfelelő terheltségű szerverek aránya.

A kiszolgálás lehet FIFO (First-In-First-Out, mint pl. a bolti sorokban), vagy párhuzamos, PS (Processor Sharing, amikor egy szerver több igényt szolgál ki párhuzamosan).

Egy szerver kiszolgálási rátája függhet attól, hogy mennyi a bent lévő igények k száma:

$$r_{(k+1) \rightarrow k}(x) = \mu_k.$$

Ez főleg PS-nél releváns, amikor egy szerver több igényt hatékonyan tud párhuzamosan kiszolgálni, mint egyet. Ennek megfelelően általában feltesszük, hogy

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq \dots \quad \text{és} \quad \mu_1 \geq \frac{\mu_2}{2} \geq \frac{\mu_3}{3} \geq \dots$$

(FIFO kiszolgálás esetén általában $\mu_1 = \mu_2 = \dots$)

A stabilitás feltétele $\lambda < \mu_K$.

Az egy sorba történő érkezési ráta viszont nem konstans. Pl. JSQ terhelés-elosztási elv esetén:

- ha van k -nál rövidebb sor, akkor k hosszú sorba nem kerül igény;
- ha k hosszú a legrövidebb sor, akkor az X_k darab k hosszú sor közül $\frac{1}{X_k}$ eséllyel kerül egy konkrét sorba az igény, így az egy adott k hosszú sorba érkező ráta

$$\frac{N\lambda}{X_k} = \frac{N\lambda}{N x_k} = \frac{\lambda}{x_k},$$

így

$$r_{k \rightarrow (k+1)}(x) = \begin{cases} 0 & \text{ha } \sum_{i=1}^{k-1} x_i > 0; \\ \frac{\lambda}{x_k} & \text{ha } \sum_{i=1}^{k-1} x_i = 0 \text{ és } x_k > 0, \end{cases}$$

(Ez egyébként nem folytonos függvény, de se baj.)

A mean-field diffegyenlet-rendszer:

$$\begin{aligned}\frac{d}{dt}v_k(t) = & r_{(k-1) \rightarrow k}(v(t))v_k(t) - r_{k \rightarrow (k+1)}(v(t))v_{k-1}(t) \\ & + \mu_{k+1}v_{k+1}(t) - \mu_k v_k(t) \quad (k = 0, \dots, K);\end{aligned}$$

a jobboldalon az egyes tagok annak felelnek meg, hogy a k hosszú sorok száma megváltozik, ha...

- érkezés történik $k - 1$ hosszú sorba, vagy
- érkezés történik k hosszú sorba, vagy
- kiszolgálás történik $k + 1$ hosszú sorban, vagy
- kiszolgálás történik k hosszú sorban.

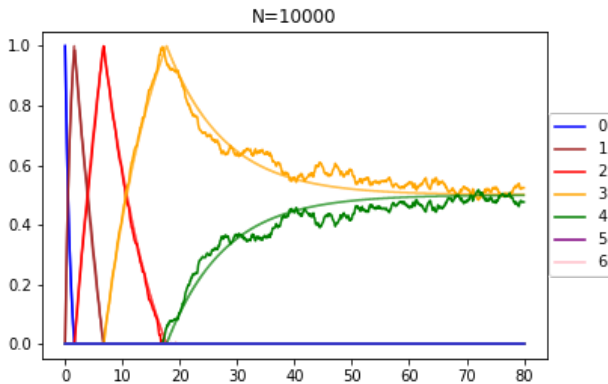
Példa. Tekintsük a következő paramétereket:

- $N = 10000$ szerver,
- egy szerver maximum 5 igényt szolgál ki egyszerre (ha több igény van bent, a többi sorban áll),
- egy szerver teljes kiszolgálási rátája a szerverben bent lévő igények számának függvényében:

k	1	2	3	4	5
μ_k	1.0	1.1	1.2	1.3	1.4

- $\lambda = 1.25$; az egész rendszerbe az érkezési ráta λN ,
- kezdeti feltétel: üres rendszer.

Hogy nézhet ki ennek a rendszer egy realizációja JSQ terheléselosztás esetén? Gondoljuk meg előre, hogy mit várunk.



Mire jó a mean-field limesz vizsgálata?

- nagyméretű rendszerek esetén a teljes rendszer állapottere hatalmas, ehelyett elég egy kisebb méretű diffegyenlet-rendszert vizsgálni, ami numerikusan könnyen megoldható;
- sok jellemző definiálható és kiszámítható a mean-field rendszerben is: például a mean-field limesz alapján megbecsülhető, hogy a populációs folyamat mikor éri el a stacionárius állapotát. De például a Little-formula is teljesül.

A korábbi példában a stabil fixpont könnyen kiszámítható.

$\lambda = 1.25$ félúton van $\mu_3 = 1.2$ és $\mu_4 = 1.3$ között, így a fixpont

$$v = (0 \ 0 \ 0 \ 0.5 \ 0.5 \ 0).$$

Alkalmazzuk a Little-formulát. Nincs adatvesztés, tehát az effektív érkezési ráta

$$\lambda_e = \lambda = 1.25,$$

az átlagos sorhossz

$$L = 0.5 \cdot 3 + 0.5 \cdot 4 = 3.5,$$

és így az átlagos kiszolgálási idő

$$W = L/\lambda = 2.8.$$

Tegyük fel, hogy adott egy eloszlás (mondjuk az $F(x)$ eloszlásfüggvénnyel). A cél, hogy generáljunk egy X valószínűségi változót ebből az eloszlásból.

Kiindulási pont: az alap véletlen forrás általában valamilyen fizikai folyamat. Pl. fotonokkal bombázunk egy érzékelőt, vagy radioaktív anyag bomlását figyeljük stb. A legegyszerűbb esetben független véletlen biteket (Bernoulli($1/2$) változók eloszlású valószínűségi változókat) tudunk generálni. Ebből akarunk bonyolultabb eloszlásokhoz eljutni.

Generálás adott eloszlásból

Az első lépés a folytonos egyenletes eloszlás.

Lemma

Ha $X_1, X_2, \dots$ független véletlen bitek, azaz

$$\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = 1/2,$$

akkor

$$U = \sum_{n=1}^{\infty} \frac{X_n}{2^n}$$

konvergens és az eloszlása $U[0, 1]$.

Megfordítva, ha $U \sim U[0, 1]$, akkor U kettes számrendszerbeli jegyei független bitek.

Biz. (vázlat) Gondoljuk meg, mi történik, ha 10-es számrendszerben a tizedesvessző után függetlenül, véletlenszerűen sorsolunk számjegyeket. A lemma ugyanez kettes számrendszerben.

A folytonos egyenletes eloszlásból a következő lemmával lehet eljutni további eloszlásokhoz.

Lemma

Ha U $[0, 1]$ -en folytonos egyenletes valószínűségi változó, akkor az

$$X = F^{-1}(U)$$

valószínűségi változó eloszlásfüggvénye F .

Generálás adott eloszlásból

Biz. U eloszlásfüggvénye

$$F_U(x) = \begin{cases} 0, & \text{ha } x < 0 \\ x, & \text{ha } 0 \leq x \leq 1 \\ 1, & \text{ha } 1 < x \end{cases}$$

Mivel F monoton növekvő, ezért $F^{-1}(U) < x$ és $U < F(x)$ ekvivalensek, és innen

$$\mathbb{P}(X < x) = \mathbb{P}(F^{-1}(U) < x) = \mathbb{P}(U < F(x)) = F_U(F(x)) = F(x).$$

Példa. Az $\text{Exp}(\lambda)$ eloszlás eloszlásfüggvénye

$$F(x) = 1 - e^{-\lambda x} \quad \rightarrow \quad F^{-1}(y) = \frac{\log(1 - y)}{\lambda},$$

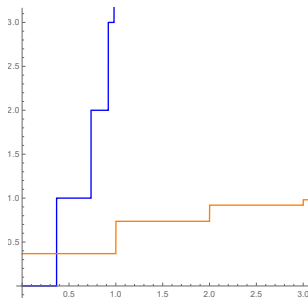
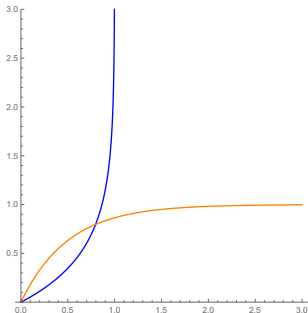
így ha $U \sim U[0, 1]$, akkor

$$X = \frac{\log(1 - U)}{\lambda} \sim \text{EXP}(\lambda).$$

Generálás adott eloszlásból

Ha F egy diszkrét eloszlás eloszlásfüggvénye, akkor vannak konstans szakaszai az egészek között. F^{-1} ilyenkor is értelmezhető, ezekben a pontokban ugrása van.

F és F^{-1} folytonos és diszkrét esetben:



Emlékeztető. Egy Markov-lánccban egy adott állapotban eltöltött idő mindig EXP eloszlású.

Láttunk már rá példát, hogy ha egy állapotot lecserélünk egy 2 állapotból álló mini Markov láncra, akkor az abban a 2 állapotban együtt eltöltött idő eloszlása már nem EXP. Ezt akarjuk tovább vizsgálni.

Általában egy Markov-lánccban helyettesíthetünk egy állapotot egy több állapotból álló mini Markov-lánccal. Ilyenkor a mini Markov-lánccnak mindig van egy nyelő állapota.

Például ha a mini Markov-lánc generátora

$$\begin{bmatrix} -1 & 1/3 & 2/3 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

ez annak felel meg, hogy amikor a nagy Markov-lánc belép a mini Markov-láncba, akkor az első és második állapot között ugrál egy ideig, majd amikor a harmadik (nyelő) állapotba lép, akkor valójában továbblép a nagy Markov-lánc egy másik állapotába.

Emiatt a harmadik állapotot nem is szokás feltüntetni:

$$A = \begin{bmatrix} -1 & 1/3 \\ 1 & -2 \end{bmatrix}$$

Definíció. Az A négyzetes mátrix szubgenerátor, ha

- A főátlóbeli elemei negatívak, főátlón kívüli elemei nemnegatívak,
- A minden egyes sorösszege legfeljebb 0, és legalább egy sorösszege szigorúan kisebb, mint 0.

Ilyenkor A tekinthető úgy, mint egy nyelővel rendelkező Markov-lánc generátora. A nyelőbe lépés rátáját az egyes állapotokból a $-A1$ oszlopvektor adja meg, ahol 1 a csupa 1-esből álló vektor.

Definíció. Legyen α eloszlás vektor és A szubgenerátor mátrix. Ekkor $\text{PH}(\alpha, A)$ az az eloszlás, ami egy α kezdeti vektorú, A szubgenerátorú, nyelővel rendelkező Markov-lánc elnyelődésének ideje.

A $\text{PH}(\alpha, A)$ eloszlás sűrűségfüggvénye

$$f(t) = \alpha e^{At}(-A1),$$

eloszlásfüggvénye

$$F(t) = 1 - \alpha e^{At}1.$$

Az A -nak megfelelő állapotokat (a mini Markov-lánc állapotait) fázisnak hívjuk.

Példa. Ha A egy 1×1 -es mátrix, akkor $\text{PH}(\alpha, A)$ az exponenciális eloszlást adja vissza.

Példa. Ha A $n \times n$ -es,

$$A = \begin{bmatrix} -\lambda & \lambda & 0 & \dots & 0 \\ 0 & -\lambda & \lambda & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & \dots & & 0 & -\lambda \end{bmatrix},$$

azaz a fázisok sorba vannak kötve, és $\alpha = (1 \ 0 \ \dots \ 0)$, akkor $\text{PH}(\alpha, A)$ éppen n darab független $\text{EXP}(\lambda)$ eloszlás összege, ami az $\text{Erlang}(\lambda, n)$ eloszlás.

PH eloszlásokkal lényegében bármilyen folytonos eloszlást lehet közelíteni.

Tétel

Tetszőleges $g(t), t \geq 0$ sűrűségfüggvényhez és $\varepsilon > 0$ számhoz létezik olyan (α, A) pár, hogy

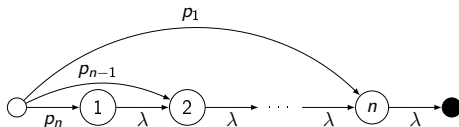
$$\int_0^{\infty} |f(t) - g(t)| < \varepsilon,$$

ahol

$$f(t) = \alpha e^{At}(-A1).$$

Általános eloszlás közelítése

Biz. (vázlat.) Koncentrált Erlang-sűrűségfüggvényekből szeretnénk kikeverni a g függvényt. Ehhez a következő struktúrát használjuk (ún. hipo-Erlang eloszlás):



a következő paraméter-választásokkal:

$$\lambda = n/T,$$
$$p_i = \frac{g(Ti/n)}{\sum_{j=1}^n g(Tj/n)}, \quad i = 1, \dots, n;$$

a p_i valószínűségű állapotból az elnyelődés ideje $\text{Erlang}(i, \lambda)$, ami az $i/\lambda = Ti/n$ pontra koncentrálódik. Ha n és T elég nagy, akkor jó lesz a közelítés.

A fenti struktúra általában nem optimális. Minél kisebb méretű az (α, A) pár, annál jobb, de kis méretű jó közelítést találni nehéz kérdés.

A PH eloszlások osztálya zárt többféle műveletre.

Tétel

Ha X és Y független PH eloszlású valószínűségi változók, akkor

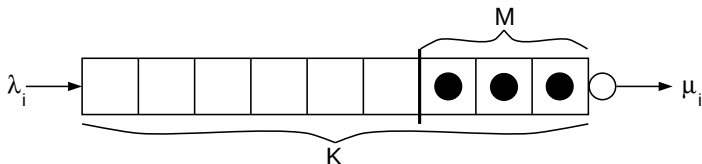
- *cX is PH eloszlású, ha $c > 0$ konstans;*
- *$X + Y$ is PH eloszlású;*
- *$\min(X, Y)$ is PH eloszlású;*
- *$\max(X, Y)$ is PH eloszlású;*
- *X és Y keveréke is PH eloszlású:*

$$Z = \begin{cases} X & p \text{ valószínűséggel} \\ Y & 1 - p \text{ valószínűséggel} \end{cases}$$

Nem biz.

Általános Markov-sor

Tekintsünk egy általános véges bufferű Markov-sort. A buffer mérete K . Az i állapotban az érkezési ráta λ_i , a kiszolgálási ráta μ_i . Feltesszük, hogy a sorhossz eloszlása stacionárius.



A kiszolgálás FIFO helyett LPS (limited processor sharing): a μ_i kiszolgálási ráta egyenletesen oszlik meg az első M igény között (vagy az összes között, ha M -nél kevesebb igény van). A további igények várakoznak.

Speciálisan az $M = 1$ esetben visszkapjuk a FIFO kiszolgálást.

Előkészület: a rendszerben töltött idő várható értéke Tekintsünk egy kitüntetett igényt egy j hosszú sorban az i -edik pozícióban ($1 \leq i \leq j \leq K$). Jelölje $X_{i,j}$ a rendszerben még hátralévő idejét, és legyen

$$W_{i,j} = E(X_{i,j}).$$

A Little-formula megadja egy véletlen igény rendszerben töltött idejének várható értékét; az összes $W_{i,j}$ ismerete ennél részletesebb információt jelent.

Megnézzük, hogyan lehet a $W_{i,j}$ -ket kiszámolni teljes várható érték tétellel.

Várható kiszolgálási idő részletesebben

A következő változások történhetnek a sorral:

- Érkezik egy igény a sor végére. Ennek rátája λ_j . Ilyenkor a kitüntetett igény továbbra is az i -edik pozícióban marad, de a sor hossza $(j + 1)$ -re változik.
- Történik egy kiszolgálás a sorban. Ennek rátája μ_j . Ha az a kitüntetett igény, akkor a kiszolgálása befejeződött. Ha egy másik igényt szolgált ki a szerver, akkor a kitüntetett igény $i > M$ esetén előrébb lép az $(i - 1)$ -edik pozícióba. $i \leq M$ esetén tulajdonképpen mindegy, előre lép-e, mivel az első M igény egyforma helyzetben van.

A kétféle változás együttes rátája $\lambda_j + \mu_j$, így a következő állapotváltozásig eltelt idő $\text{EXP}(\lambda_j + \mu_j)$, aminek a várható értéke $\frac{1}{\lambda_j + \mu_j}$, és a változáskor

- $\frac{\lambda_j}{\lambda_j + \mu_j}$ az érkezés valószínűsége, és
- $\frac{\mu_j}{\lambda_j + \mu_j}$ a kiszolgálás valószínűsége.

Várható kiszolgálási idő részletesebben

Legyen először $M < i \leq j < K$. Ilyenkor a kitüntetett igény nem kap kiszolgálást, hanem csak várakozik; a rendszerben hátralévő idejének várható értéke $W_{i,j}$. A teljes várható érték tétel alapján

$$W_{i,j} = \frac{1}{\lambda_j + \mu_j} + \left(\frac{\lambda_j}{\lambda_j + \mu_j} W_{i,j+1} + \frac{\mu_j}{\lambda_j + \mu_j} W_{i-1,j-1} \right),$$

ahol $\frac{1}{\lambda_j + \mu_j}$ az első változásig eltelt idő várható értéke, és ...

- ha érkezés történt, akkor a kitüntetett igény marad az i -edik helyen, de a sor hossza $j + 1$ -re nőtt, és onnan a kiszolgálásig szükséges további idő várható értéke $W_{i,j+1}$;
- ha kiszolgálás történt, akkor a kitüntetett igény pozíciója és a sor hossza is csökken 1-gyel, és a kiszolgálásig szükséges további idő várható értéke $W_{i-1,j-1}$.

A cél: felírni hasonló egyenleteket az összes $W_{i,j}$ -re, majd a kapott lineáris egyenletrendszert megoldani.

További esetek.

Ha $i \leq j \leq M$ és $j < K$, akkor egyrészt

$$W_{1,j} = \dots = W_{j,j},$$

hiszen az első j igény helyzete egyforma, másrészt

$$W_{i,j} = \frac{1}{\lambda_j + \mu_j} + \\ + \left(\frac{\lambda_j}{\lambda_j + \mu_j} W_{i,j+1} + \frac{\mu_j}{\lambda_j + \mu_j} \frac{j-1}{j} W_{i-1,j-1} + \frac{\mu_j}{\lambda_j + \mu_j} \frac{1}{j} \cdot 0 \right),$$

mivel az első j igényből $1/j$ eséllyel a kitüntetett igényt szolgálja ki a szerver, és olyankor az elhagyja a rendszert (0 további időt tölt a rendszerben).

Ha $i \leq M < j < K$, akkor is

$$W_{1,j} = \dots = W_{M,j},$$

hiszen az első M igény helyzete egyforma, másrészt

$$W_{i,j} = \frac{1}{\lambda_j + \mu_j} + \left(\frac{\lambda_j}{\lambda_j + \mu_j} W_{i,j+1} + \frac{\mu_j}{\lambda_j + \mu_j} \frac{M-1}{M} W_{i-1,j-1} + \frac{\mu_j}{\lambda_j + \mu_j} \frac{1}{M} \cdot 0 \right),$$

mivel ilyenkor az első M igény között oszlik meg a kiszolgálási ráta, és azon belül $1/M$ eséllyel a kitüntetett igényt szolgálja ki a szerver.

$j = K$, azaz tele buffer esetén nem történhet érkezés, csak kiszolgálás. Ha $i \leq M$, akkor is teljesül

$$W_{1,K} = \dots = W_{M,K},$$

továbbá

$$W_{i,K} = \frac{1}{\mu_j} + \left(\frac{M-1}{M} W_{i-1,M-1} + \frac{1}{M} \cdot 0 \right),$$

Hasonlóan tele buffer, csak $i > M$ esetén a kitüntetett igény biztosan nem kap kiszolgálást, így

$$W_{i,K} = \frac{1}{\mu_j} + W_{i-1,M-1}.$$

Várható kiszolgálási idő részletesebben

A fentiek megadnak a $W_{i,j}$ ismeretlenekre egy lineáris egyenletrendszert, amit meg lehet oldani.

A $W_{i,j}$ -k ismeretében megadható egy véletlen igény rendszerben töltött idejének várható értéke, W is (amit egyébként a Little-formulából is ki tudunk számolni). Arra kell figyelniük, hogy általános Markov-sorra az érkezési ráta függ az állapottól.

Ha a sorhossz stacionárius eloszlása $v_{st} = (x_0 \ x_1 \ \dots \ x_K)$, akkor az effektív érkezési ráta

$$\lambda_e = x_0 \lambda_0 + x_1 \lambda_1 + \dots + x_{K-1} \lambda_{K-1},$$

$$P(\text{egy véletlen igény } i \text{ hosszú sorba érkezik}) = \frac{x_i \lambda_i}{\lambda_e},$$

végül

$$W = \frac{x_0 \lambda_0}{\lambda_e} W_{1,1} + \frac{x_1 \lambda_1}{\lambda_e} W_{2,2} + \dots + \frac{x_{K-1} \lambda_{K-1}}{\lambda_e} W_{K,K}.$$

Előkészület vége, a következő már a teljes eloszlás kiszámítása Laplace-transzformálttal.

Tekintsünk egy kitüntetett igényt egy j hosszú sorban az i -edik pozícióban ($1 \leq i \leq j \leq K$). $X_{i,j}$ a kiszolgálási ideje.

Legyen $X_{i,j}$ sűrűségfüggvénye $f_{i,j}(t)$, Laplace-transzformáltja $f_{i,j}^*(s)$.

A cél: a korábbi $W_{i,j}$ -khez hasonló egyenleteket felírni, csak ezúttal $f_{i,j}^*(s)$ -re, majd a kapott egyenletrendszert megoldani.

A következő változások történhetnek a sorral:

- Érkezik egy igény a sor végére λ_j rátával, vagy
- történik egy kiszolgálás a sorban μ_j rátával.

A kétféle változás együttes rátája $\lambda_j + \mu_j$, így a következő állapotváltozásig eltelt idő $\text{EXP}(\lambda_j + \mu_j)$ eloszlású, ennek Laplace-transzformáltja $\frac{\lambda_j + \mu_j}{\lambda_j + \mu_j + s}$.

A változáskor $\frac{\lambda_j}{\lambda_j + \mu_j}$ az érkezés valószínűsége és $\frac{\mu_j}{\lambda_j + \mu_j}$ a kiszolgálás valószínűsége.

Legyen először $M < i \leq j < K$. A fentiek alapján a Laplace-transzformáltra teljesül a következő egyenlet:

$$f_{i,j}^*(s) = \frac{\lambda_j + \mu_j}{\lambda_j + \mu_j + s} \left(\frac{\lambda_j}{\lambda_j + \mu_j} f_{i,j+1}^*(s) + \frac{\mu_j}{\lambda_j + \mu_j} f_{i-1,j-1}^*(s) \right)$$

További esetek.

Ha $i \leq j \leq M$ és $j < K$, akkor egyrészt

$$f_{1,j}^*(s) = \cdots = f_{j,j}^*(s),$$

hiszen az első j igény helyzete egyforma, másrészt

$$f_{i,j}^*(s) = \frac{\lambda_j + \mu_j}{\lambda_j + \mu_j + s} \times \\ \left(\frac{\lambda_j}{\lambda_j + \mu_j} f_{i,j+1}^*(s) + \frac{\mu_j}{\lambda_j + \mu_j} \frac{j-1}{j} f_{i-1,j-1}^*(s) + \frac{\mu_j}{\lambda_j + \mu_j} \frac{1}{j} \cdot 1 \right),$$

mivel az első j igényből $1/j$ eséllyel a kitüntetett igényt szolgálja ki a szerver, és olyankor az elhagyja a rendszert (0 további időt tölt a rendszerben, aminek a Laplace-transzformáltja 1).

Ha $i \leq M < j < K$, akkor is

$$f_{1,j}^*(s) = \dots = f_{M,j}^*(s),$$

hiszen az első M igény helyzete egyforma, másrészt

$$f_{i,j}^*(s) = \frac{\lambda_j + \mu_j}{\lambda_j + \mu_j + s} \times \\ \left(\frac{\lambda_j}{\lambda_j + \mu_j} f_{i,j+1}^*(s) + \frac{\mu_j}{\lambda_j + \mu_j} \frac{M-1}{M} f_{i-1,j-1}^*(s) + \frac{\mu_j}{\lambda_j + \mu_j} \frac{1}{M} \cdot 1 \right),$$

mivel ilyenkor az első M igény között oszlik meg a kiszolgálási ráta, és azon belül $1/M$ eséllyel a kitüntetett igényt szolgálja ki a szerver.

$j = K$, azaz tele buffer esetén nem történhet érkezés, csak kiszolgálás. Ha $i \leq M$, akkor is teljesül

$$f_{1,K}^*(s) = \dots = f_{M,K}^*(s),$$

továbbá

$$f_{i,K}^*(s) = \frac{\mu_j}{\mu_j + s} \left(\frac{M-1}{M} f_{i-1,M-1}^*(s) + \frac{1}{M} \cdot 1 \right),$$

Hasonlóan tele buffer, csak $i > M$ esetén a kitüntetett igény biztosan nem kap kiszolgálást, így

$$f_{i,K}^*(s) = \frac{\mu_j}{\mu_j + s} f_{i-1,M-1}^*(s).$$

Az előbbiek megadnak az ismeretlen $f_{i,j}^*(s)$ függvényekre egy lineáris egyenletrendszert, amiben az együtthatók ismert számok és függvények, így az egyenletrendszert meg tudjuk oldani.

Az $f_{i,j}^*(s)$ függvények racionális törtfüggvények lesznek, azaz $\frac{p(s)}{q(s)}$ alakúak, ahol $p(s)$ és $q(s)$ polinomok.

Az $f_{i,j}^*(s)$ -ek ismeretében megadható egy véletlen igény kiszolgálási idejének Laplace-transzformáltja, $f^*(s)$ is, ugyanolyan súlyozással, mint korábban:

$$f^*(s) = \frac{x_0 \lambda_0}{\lambda_e} f_{1,1}^*(s) + \frac{x_1 \lambda_1}{\lambda_e} f_{2,2}^*(s) + \cdots + \frac{x_{K-1} \lambda_{K-1}}{\lambda_e} f_{K,K}^*(s).$$

Végül annyi van hátra, hogy $f^*(s)$ -et inverz Laplace transzformáljuk (ILT), és megkapjuk a kiszolgálási idő eloszlásának az $f(t)$ sűrűségfüggvényét. Ha az $F(t)$ eloszlásfüggvényre van szükségünk, akkor annyi a teendő, hogy $f^*(s)/s$ -et inverz Laplace-transzformáljuk.

$f^*(s)$ is racionális törtfüggvény, és a racionális törtfüggvényeknek az inverz Laplace-transzformáltja analitikusan is kiszámítható parciális törtekre bontással.

Még egy apróság: ugyan $f(t)$ elvileg több információt tartalmaz, mint a várható érték (és $f(t)$ -ből ki is tudnánk számítani integrálással), de ha csak a várható értékre van szükségünk, akkor közvetlenül a várható értéket kiszámítani általában gyorsabb és numerikusan stabilabb.