Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Mikael Raihhelgauz

Dirichlet' protsess

Matemaatika eriala
Bakalaureusetöö (9 EAP)

Juhendaja: professor Jüri Lember

Tartu 2020

Dirichlet' protsess

Bakalaureusetöö

Mikael Raihhelgauz

Lühikokkuvõte. Käesolevas bakalaureusetöös antakse referatiivne ülevaade Dirichlet' protsessi tähtsamatest omadustest ning selgitatakse nende olulisust mitteparameetrilise Bayesi statistika raames. Samuti käsitletakse Dirichlet' protsessi konstrueerimise mõõduteoreetilisi eeldusi ning lõplikumõõtmelist Dirichlet' jaotust. Töö tugineb suuresti Patrick Billingsley raamatule *"Weak Convergence of Measures"* ja Subhashis Ghosali ning Aad van der Vaarti raamatule *"Fundamentals of Nonparametric Bayesian Inference"*.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad. Juhuslikud protsessid, Dirichlet' jaotus, Dirichlet' protsess, mõõduteooria.

Dirichlet process

Bachelor's thesis

Mikael Raihhelgauz

Abstract. The objective of this bachelor's thesis is to present an overview of the fundamental properties of the Dirichlet process and explain their significance within the framework of Bayesian nonparametric inference. The measure-theoretic preconditions for constructing the Dirichlet process and general properties of the finite-dimensional Dirichlet distribution are assessed as well. The thesis is primarily based on Patrick Billingsley's *"Weak Convergence of Measures"* and *"Fundamentals of Nonparametric Bayesian Inference"* by Subhashis Ghosal and Aad van der Vaart.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics.

Keywords. Random processes, Dirichlet distribution, Dirichlet process, measure theory.

Sisukord

Sissejuhatus

Paljude statistiliste probleemide korral puuduvad täpsed teadmised vaatlusi genereerinud jaotusest p. Seega mitteparameetrilise Bayesi lähenemisviisi puhul vaadeldakse jaotust juhusliku suurusena P. Esialgset arusaama P jaotusest postuleeritakse eeljaotusena. Vaatluste sissearvestamisel seda arusaama uuendatakse ja saadakse järeljaotus. Üldjuhul on eeljaotuse valikul kaks miinimumnõuet. Esiteks, peab järeljaotus kuuluma eeljaotusega samasse jaotuste perre. Teiseks, peab P järeljaotus koonduma vaatluste arvu kasvades tõeliseks jaotuseks p. 1973. aastal Fergusoni kirjeldatud Dirichlet' protsessi jaotus [3] rahuldab mõlemat omadust. Tänaseks on sellest kujunenud levinuim eeljaotuse valik mitteparameetriliste probleemide puhul [4, lk 35–36].

Käesoleva bakalaureusetöö eesmärk on anda referatiivne ülevaade mõõduteoreetilisest baasist, millele tugineb Dirichlet' protsessi konstruktsioon ja tutvustada protsessi tähtsamaid omadusi. Lisaks sellele selgitame Dirichlet' protsessi tähtsust töövahendina mitteparameetrilise Bayesi statistika raamistikus.

Esimseses peatükis esitame Dirichlet' protsessi defineerimiseks vajaliku matemaatilise aparatuuri. Vaatleme kõigi meetrilisel ruumil määratud tõenäosusmõõtude hulka $\mathcal{P}$ ja esitame viisi vastava tõenäosusruumi $(\mathcal{P}, \mathcal{B}(\mathcal{P}), \Pi)$ konstrueerimiseks.

Teises peatükis defineerime lõplikumõõtmelise Dirichlet' jaotuse ühiksimpleksil, millele omakorda tugineb Dirichlet' protsessi definitsioon. Seejärel anname ülevaate jaotuse tähtsamatest omadustest ning esitame lihtsa skeemi Dirichlet' jaotusega juhusliku vektori genereerimiseks.

Kolmandas peatükis anname Dirichlet' protsessi definitsiooni, vaatleme Dirichlet' protsessi olemasoluga seotud küsimusi ja uurime selle olulisimaid omadusi. Selgitame ka Dirichlet' protsessi rakendusvõimalusi mitteparameetrilises Bayesi statistikas ning lõpuks kirjeldame lihtsaid algoritme Dirichlet' protsessi jaotusega juhusliku mõõdu ja Dirichlet' protsessist juhusliku valimi genereerimiseks.

1. Tõenäosusruum $(\mathcal{P}, \mathcal{B}(\mathcal{P}), \Pi)$

Olgu (S, d) meetriline ruum, $\mathcal{B}(S)$ vastav Boreli σ-algebra ja $\mathcal{P}$ kõigi mõõtuval ruumil $(S, \mathcal{B}(S))$ määratud tõenäosusmõõtude hulk. Selles peatükis esitame vajaliku matemaatilise aparatuuri defineerimaks tõenäosusmõõtu ruumil $\mathcal{P}$. Märgime etteruttavalt, et Dirichlet' protsessi jaotus on niisugune tõenäosusmõõt, seega range mõõduteoreetiline baas on tarvilik protsessi korrektseks konstrueerimiseks ja selle omaduste kirjeldamiseks. Antud peatüki materjal tugineb põhiliselt [1, ptk 5–6].

Definitsioon 1.1. Tõenäosusmõõtude nõrk koondumine. Öeldakse, et tõenäosusmõõdud $p_n \in \mathcal{P}$, $n \in \mathbb{N}$, koonduvad nõrgalt mõõduks $p \in \mathcal{P}$ ja kirjutatakse $p_n \Rightarrow p$, kui iga pideva tõkestatud funktsiooni $f : S \to \mathbb{R}$ korral

$$\lim_{n\to\infty} \int_S f dp_n = \int_S f dp.$$

Järgnev teoreem loetleb nõrga koondumisega samaväärseid tingimusi, luues edasise teooriaarenduse seisukohalt mugava töövahendi. Selle tõestus on esitatud näiteks [1, lk 16–17].

Teoreem 1.2. *Portmanteau' teoreem.* *Olgu p_n, $n \in \mathbb{N}$, ja p tõenäosusmõõdud ruumil $(S, \mathcal{B}(S))$. Järgmised tingimused on ekvivalentsed:*

1. $p_n \Rightarrow p$;
2. *iga ühtlaselt pideva tõkestatud funktsiooni $f : S \to \mathbb{R}$ korral*
$$\lim_{n\to\infty} \int_S f dp_n = \int_S f \, dp;$$
3. *iga kinnise hulga $F \in \mathcal{B}(S)$ korral $\limsup_n p_n(F) \leq p(F)$;*
4. *iga lahtise hulga $G \in \mathcal{B}(S)$ korral $\liminf_n p_n(G) \geq p(G)$;*
5. *kui $A \in \mathcal{B}(S)$ korral $p(\partial A) = 0$, siis $\lim_n p_n(A) = p(A)$.*

Järgmise abilemma tõestuse leiab [1, lk 7–8].

Lemma 1.3. *Tõenäosusmõõt $p \in \mathcal{P}$ on üheselt määratud väärtustega $p(F)$ kinnistel hulkadel $F \in \mathcal{B}(S)$.*

Defineerimaks tõenäosusmõõtu hulgal $\mathcal{P}$ peame kõigepealt määratlema, millised $\mathcal{P}$ alamhulgad on mõõtuvad. Selleks vajame vastavat Boreli σ-algebrat $\mathcal{B}(\mathcal{P})$ ehk vähimat hulga $\mathcal{P}$ lahtiseid alamhulki sisaldavat σ-algebrat. Siin aga peame omakorda täpsustama, milliseid hulki peame lahtisteks. Antud küsimusele vastamiseks toome sisse Prohhorovi meetrika mõiste ja veendume, et tegemist on kaugusega. Edaspidi hulga $\mathcal{P}$ lahtistest hulkadest rääkides, peame silmas just Prohhorovi meetrika indutseeritud topoloogia elemente.

Definitsioon 1.4. Prohhorovi meetrika. Prohhorovi meetrikaks nimetatakse funktsiooni $\rho : \mathcal{P} \times \mathcal{P} \to [0, \infty)$, mis on defineeritud järgneva seosega:

$$\rho(p, q) = \inf\{\epsilon > 0 : \text{ iga Boreli hulga } A \text{ korral } p(A) \le q(A^\epsilon) + \epsilon \text{ ja } q(A) \le p(A^\epsilon) + \epsilon\},$$

kus $A^\epsilon = \{x \in S : d(x, A) < \epsilon\}$.

Lause 1.5. *Prohhorovi meetrika on kaugus.*

Tõestus. **M1.** Definitsioonist on ilmne, et ρ on mittenegatiivne funktsioon.

M2. Fikseerime vabalt $p, q \in \mathcal{P}$. Siis

$$\rho(p, q) = \inf\{\epsilon > 0 : \forall A \in \mathcal{B}(S)\ p(A) \le q(A^\epsilon) + \epsilon \ \&\ q(A) \le p(A^\epsilon) + \epsilon\} = \rho(q, p).$$

M3. Olgu $p \in \mathcal{P}$. Siis $\rho(p, p) = \inf\{\epsilon > 0 : \text{ iga Boreli hulga } A \text{ korral } p(A) \le p(A^\epsilon) + \epsilon\}$. Et $A \subset A^\epsilon$ iga $\epsilon > 0$ korral, siis ka $p(A) \le p(A^\epsilon)$ iga $\epsilon > 0$ korral. Järelikult $\rho(p, p) = 0$. Nüüd eeldame, et $\rho(p, q) = 0$. See tähendab, et suvalise $\epsilon > 0$ korral $p(A) \le q(A^\epsilon) + \epsilon$. Olgu A kinnine. Vaatleme jada $A_n = A^{\frac{1}{n}}, n \in \mathbb{N}$. On selge, et $A_k \subset A_m$, kui $k \ge m$. Paneme tähele, et $A = \bigcap_{n=1}^{\infty} A_n$. Tõepoolest, kuna iga $n \in \mathbb{N}$ korral $A \subset A^{\frac{1}{n}}$, siis $A \subset \bigcap_{n=1}^{\infty} A_n$. Kehtib ka vastupidine sisalduvus. Kui $x \in \bigcap_{n=1}^{\infty} A_n$, siis $d(x, A) < \dfrac{1}{n}$ iga $n \in \mathbb{N}$ korral ehk $\inf\{d(x, a) : a \in A\} = 0$. See tähendab, et $x \in A$ või $x \in \partial A$. Kuna A on kinnine, siis $\partial A \subset A$, seega x on igal juhul hulga A element. Niisiis, $\bigcap_{n=1}^{\infty} A_n \subset A$. Järelikult $\lim_n q(A^{\frac{1}{n}}) = q\Big(\bigcap_{n=1}^{\infty} A_n\Big) = q(A)$. Piirväärtuse monotoonsuse tõttu $p(A) \le q(A)$

Samuti kehtib võrratus $q(A) \le p(A^\epsilon) + \epsilon$. Juhul, kui $\epsilon \downarrow 0$, saame piirväärtuse monotoonsuse tõttu, et kinnise A korral kehtib ka $q(A) \le p(A)$. Kokkuvõttes oleme saanud, et iga kinnise $A \in \mathcal{B}(S)$ puhul $p(A) = q(A)$. Lemma 1.3 põhjal $p = q$.

M4. Fikseerime vabalt $p, q, r \in \mathcal{P}$. Kui $\rho(p,q) < \epsilon_1$ ja $\rho(q,r) < \epsilon_2$, siis

$$p(A) \le q(A^{\epsilon_1}) + \epsilon_1 \le r((A^{\epsilon_1})^{\epsilon_2}) + \epsilon_1 + \epsilon_2 \le r(A^{\epsilon_1+\epsilon_2}) + \epsilon_1 + \epsilon_2,$$
$$r(A) \le q(A^{\epsilon_2}) + \epsilon_2 \le p((A^{\epsilon_1})^{\epsilon_2}) + \epsilon_1 + \epsilon_2 \le p(A^{\epsilon_1+\epsilon_2}) + \epsilon_1 + \epsilon_2.$$

Järelikult $\rho(p,r) \le \epsilon_1 + \epsilon_2$ ning ühtlasi

$$\rho(p,r) \le \inf\{\epsilon_1 + \epsilon_2 : \rho(p,q) < \epsilon_1, \rho(q,r) < \epsilon_2\} =$$
$$= \inf\{\epsilon_1 : \rho(p,q) < \epsilon_1\} + \inf\{\epsilon_2 : \rho(q,r) < \epsilon_2\} = \rho(p,q) + \rho(q,r).$$

□

Järgnevalt avame Prohhorovi meetrika seost nõrga koondumisega. Nimelt osutub, et tõenäosusmõõtude nõrk koondumine järeldub vahetult Prohhorovi meetrika järgi koondumisest. Sepraabli meetrilise ruumi S korral on aga nõrk ja meetrika järgi koondumine ekvivalentsed.

Lemma 1.6. *Olgu $p, q \in \mathcal{P}$. Kui $p(A) \le q(A^\epsilon) + \epsilon$ iga $A \in \mathcal{B}(S)$ korral, siis $\rho(p,\ q)\ \le\ \epsilon$.*

Tõestus. Sisuliselt tahame näidata, et juhul, kui $p(A) \le q(A^\epsilon) + \epsilon$ iga $A \in \mathcal{B}(S)$ korral, siis iga Boreli hulga puhul kehtib ka võrratus $q(A) \le p(A^\epsilon) + \epsilon$. Paneme tähele, et sisalduvused $A \subset S \setminus B^\epsilon$ ja $B \subset S \setminus A^\epsilon$ on ekvivalentsed. Tõepoolest,

$$A \subset S \setminus B^\epsilon \Leftrightarrow$$
$$\Leftrightarrow a \in A \Rightarrow a \in S \wedge a \notin B^\epsilon \Leftrightarrow$$
$$\Leftrightarrow a \in A \Rightarrow a \in S \wedge a \notin \{x \in S : d(x,B) < \epsilon\} \Leftrightarrow$$
$$\Leftrightarrow \text{ suvaliste } a \in A \text{ ja } b \in B \text{ puhul } d(a,b) \ge \epsilon \Leftrightarrow$$
$$\Leftrightarrow b \in B \Rightarrow b \in \{x \in S : d(x,A) \ge \epsilon\} \Leftrightarrow$$
$$\Leftrightarrow b \in B \Rightarrow b \in S \setminus A^\epsilon \Leftrightarrow$$
$$\Leftrightarrow B \subset S \setminus A^\epsilon$$

Eeldame, et iga Boreli hulga A korral kehtib $p(A) \leq q(A^\epsilon) + \epsilon$. Tähistame $B = S \setminus A^\epsilon$. Siis

$$p(A^\epsilon) = 1 - p(B) \geq 1 - q(B^\epsilon) - \epsilon = q(S \setminus B^\epsilon) - \epsilon \geq q(A) - \epsilon,$$
$$p(A^\epsilon) + \epsilon \geq q(A).$$

□

Lause 1.7. *Olgu $p_n, p \in \mathcal{P}$, $n \in \mathbb{N}$, sellised, et $\rho(p_n, p) \xrightarrow{n} 0$. Siis $p_n \Rightarrow p$.*

Tõestus. Fikseerime $p \in \mathcal{P}$ ja $p_n \in \mathcal{P}$, $n \in \mathbb{N}$, nii, et $\rho(p_n, p) \xrightarrow{n} 0$. Vaatleme kahanevat jada (ϵ_n), kus $\epsilon_n := \min_{k \leq n} \rho(p_k, p) + \frac{1}{n}$. Suvalise kinnise hulga $F \in \mathcal{B}(S)$ korral kehtib võrratus $p_n(F) \leq p(F^{\epsilon_n}) + \epsilon_n$ ja $\epsilon_n \xrightarrow{n} 0$. Ühtlasi $F^{\epsilon_k} \subset F^{\epsilon_m}$, kui $k \geq m$, ja $\bigcap_{n=1}^{\infty} F^{\epsilon_n} = F$. Järelikult

$$\limsup_n p_n(F) \leq \limsup_n \big(p(F^{\epsilon_n}) + \epsilon_n\big) = p(F).$$

Portmanteau' teoreemi kohaselt $p_n \Rightarrow p$. □

Lause 1.8. *Kui S on separaabel ja $p_n \Rightarrow p$, siis $\rho(p_n, p) \xrightarrow{n} 0$.*

Tõestus. Fikseerime vabalt $\epsilon > 0$. Olgu $\{A_i\}$ ruumi S selline mõõtuv tükeldus, kus iga hulga diameeter on väiksem kui ϵ. Märgime, et separaabluse tõttu selline tükeldus leidub. Näiteks võime võtta kogumi $\{B(x, \frac{\epsilon}{2}) : x \in D\}$, kus $D \subset S$ on ruumi S mingi kõikjal tihe ülimalt loenduv alamhulk. Määrates igale hulga D elemendile x mingi indeksi $i \in \mathbb{N}$, moodustame lõikumatud hulgad $A_1 := B(x_1, \frac{\epsilon}{2})$, $A_i := B(x_i, \frac{\epsilon}{2}) \setminus \bigcup_{j=1}^{i-1} A_j$, kui $i \geq 2$. Nii saamegi sobiva tükelduse.
Valime $k \in \mathbb{N}$ nii, et $p(\bigcup_{i>k} A_i) < \epsilon$. See on võimalik, sest $\lim_n p(\bigcup_{i=1}^{n} A_i) = p(S) = 1$ ning vastavalt $\lim_n p(\bigcup_{i>n} A_i) = \lim_n p(S \setminus \bigcup_{i=1}^{n} A_i) = \lim_n \big(1 - p(\bigcup_{i=1}^{n} A_i)\big) = 0$.
Vaatleme lõplikku lahtiste hulkade kogumit

$$\mathcal{G} = \{(A_{i_1} \cup \ldots \cup A_{i_m})^\epsilon : 1 \leq i_1 < \ldots < i_m \leq k\}.$$

Olgu $A \in \mathcal{B}(S)$ suvaline ja

$$A_0 := \bigcup_{\substack{1 \leq i \leq k, \\ A_i \cap A \neq \varnothing}} A_i.$$

Selge, et $A_0^\epsilon \in \mathcal{G}$. Kuna $p_n \Rightarrow p$, siis Portmanteau' teoreemi kohaselt $\liminf_n p_n(G) \geq p(G)$ iga lahtise hulga G korral. See tähendeab, et ka suvliase $G \in \mathcal{G}$ puhul leidub N_G nii, et kehtib implikatsioon $n \geq N_G \Rightarrow p_n(G) > p(G) - \epsilon$. Kui $n \geq N_{A_0^\epsilon}$, siis

$$p(A) \leq p(A_0) + p(\bigcup_{i>k} A_i) \leq p(A_0) + \epsilon \leq p(A_0^\epsilon) + \epsilon < p_n(A_0^\epsilon) + 2\epsilon.$$

Paneme tähele, et $A_0^\epsilon \subset A^{2\epsilon}$. Tõepoolest, kui $a \in A_0^\epsilon$, siis leidub

$$j \in \left\{n \in \{1, \dots, k\} : A \cap A_n \neq \varnothing\right\}$$

ja $x \in A_j$ nii, et $d(a,x) < \epsilon$. Fikseerime $y \in A \cap A_j$. Siis $d(x,y) \leq \epsilon$. Kolmunrga võrratuse tõttu $d(a,y) \leq d(a,x) + d(x,y) < 2\epsilon$, seega $a \in A^{2\epsilon}$. Niisiis,

$$p(A) \leq p_n(A_0^\epsilon) + 2\epsilon \leq p_n(A^{2\epsilon}) + 2\epsilon.$$

Lemma 1.6 põhjal $\rho(p_n, p) \leq 2\epsilon$. Järelikult $\rho(p_n, p) \xrightarrow{n} 0$. □

Järgmine teoreem väidab, et separaablil meetrilisel ruumil määratud tõenäosusmõõtude ruum on samuti separaabel. Eriti märkimisväärne on see, et tõestuses konstrueeritud ruumi $\mathcal{P}$ kõikjal tihe alamhulk sisaldab ainult diskreetseid mõõte. Niisiis, suvalise tõenäosusmõõdu $p \in \mathcal{P}$ korral leidub (Prohhorovi meetrika mõttes) kuitahes lähedal asuvaid diskreetseid mõõte. Praktikas tähendab see, et saame lähendada pidevaid jaotusi separaablil meetrilisel ruumil (nt $\mathbb{R}$ või $\mathbb{R}^n$) kuitahes täpselt diskreetsete jaotustega.

Teoreem 1.9. *Kui S on separaabel, siis ka $\mathcal{P}$ on separaabel.*

Tõestus. Fikseerime vabalt $\epsilon > 0$. Eeldame, et S on separaabel ja valime S tükelduse $\{A_i\}$ nii, et iga $i \in \mathbb{N}$ korral $\operatorname{diam} A_i < \epsilon$ (üks võimalik protseduur tükelduse konstrueerimiseks on kirjeldatud eelneva teoreemi tõestuses). Igast mittetühjast hulgast A_i valime elemendi x_i. Vaatleme kogumit

$$\Pi = \left\{ \sum_{i=1}^k r_i \delta_{x_i} : k \in \mathbb{N},\ \sum_{i=1}^k r_i = 1,\ r_i \in \mathbb{Q} \cap [0, \infty) \right\}.$$

Suvalise $p \in \mathcal{P}$ korral valime $k \in \mathbb{N}$ nii, et $p(\bigcup_{i>k} A_i) < \epsilon$. Nüüd valime $r_1, \dots, r_k \in \mathbb{Q} \cap [0, \infty)$ nii, et $\sum_{i=1}^k r_i = 1$ ja $\sum_{i=1}^k |r_i - p(A_i)| < \epsilon$. Veendume, et see on võimalik. Märgime, et

$$\sum_{i=1}^k p(A_i) = p(\bigcup_{i=1}^k A_i) = 1 - p(\bigcup_{i>k} A_i) > 1 - \epsilon$$

ehk $\epsilon > 1 - \sum_{i=1}^{k} p(A_i)$. Tähistame $\delta := \epsilon - 1 + \sum_{i=1}^{k} p(A_i)$. Kui $i \leq k-1$ valime mittenegatiivsed ratsionaalarvud r_i nii, et $p(A_i) \leq r_i < p(A_i) + \frac{\delta}{k}$, ja olgu $r_k = 1 - \sum_{i=1}^{k-1} r_i$. Sellisel juhul

$$\sum_{i=1}^{k} |r_i - p(A_i)| = \sum_{i=1}^{k} (r_i - p(A_i)) < \frac{(k-1)\delta}{k} + r_k - p(A_k)$$

$$= \frac{(k-1)\delta}{k} + 1 - \sum_{i=1}^{k-1} r_i - p(A_k) \leq \frac{(k-1)\delta}{k} + 1 - \sum_{i=1}^{k} p(A_i) \leq \delta + 1 - \sum_{i=1}^{k} p(A_i) = \epsilon.$$

Nüüd vaatleme tõenäosusmõõtu $q = \sum_{i \leq k} r_i \delta_{x_i}$. Suvalise hulga A korral olgu I kõigi selliste indeksite $i \leq k$ hulk, mille korral $A \cap A_i \neq \varnothing$. Defineerime hulga $A_0 := \bigcup_{i \in I} A_i$. Siis

$$p(A) \leq p(A_0) + p(\bigcup_{i>k} A_i) < p(A_0) + \epsilon = \sum_{i \in I} p(A_i) + \epsilon \leq \sum_{i \in I} r_i + 2\epsilon = q(A_0) + 2\epsilon \leq q(A^{2\epsilon}) + 2\epsilon.$$

Lemma 1.6 põhjal tähendab see, et $\rho(p, q) < 2\epsilon$. Järelikult iga $p \in \mathcal{P}$ korral võime leida kuitahes lähedal paikneva mõõdu $q \in \Pi$. Niisiis, tegemist on kõikjal tiheda ülimalt loenduva alamhulgaga. $\square$

Nüüd saame tõestada veelgi tugevama seose ruumi S ja $\mathcal{P}$ vahel. Osutub, et juhul, kui S on Poola ruum – s.t separaabel täielik meetriline ruum [1] –, siis ka $\mathcal{P}$ on Poola ruum. Selle tõestuse tarbeks toome sisse tõenäosusmõõtude kogumi suhtelise kompaktsuse ja tiheduse mõisted ning esitame Prohhorovi teoreemi.

Definitsioon 1.10. Tõenäosusmõõtude kogumi suhteline kompaktsus. Olgu $\mathfrak{P} \subset \mathcal{P}$. Öeldakse, et tõenäosusmõõtude kogum $\mathfrak{P}$ on suhteliselt kompaktne, kui igast $\mathfrak{P}$ elementide jadast saab eraldada kauguse ρ järgi koonduva osajada.

Definitsioon 1.11. Tõenäosusmõõtude kogumi tihedus. Öeldakse, et tõenäosusmõõtude kogum $\mathfrak{P} \subset \mathcal{P}$ on tihe, kui iga $\epsilon > 0$ korral leidub kompaktne hulk $K \in \mathcal{B}(S)$ nii, et iga $p \in \mathfrak{P}$ korral $p(K) > 1 - \epsilon$.

[1] Formaalselt nimetatakse Poola ruumiks sepraablit täielikult metriseeritavat *topoloogilist* ruumi. Meie käsitluses on aga meetrika d fikseeritud, seega öeldes, et S on Poola ruum, peame silmas seda, et meetriline ruum (S, d) on separaabel ja täielik.

Teoreem 1.12. ***Prohhorovi teoreem.*** *Olgu $\mathfrak{P} \subset \mathcal{P}$. Kui $\mathfrak{P}$ on tihe, siis see on suhteliselt kompaktne. Olgu S separaabel ja täielik. Siis $\mathfrak{P}$ on tihe, kui see on suhteliselt kompaktne.*

Prohhorovi teoreemi tõestuse leiab näiteks [1, lk 59–63]

Definitsioon 1.13. Täielikult tõkestatud hulk. Öeldakse, et hulk $A \subset S$ on täielikult tõkestatud, kui iga $\epsilon > 0$ korral leidub arv $n \in \mathbb{N}$ ning hulgad $A_1, \dots, A_n \subset S$ nii, et $\operatorname{diam}(A_i) < \epsilon$, $i = 1, \dots, n$, ja $A \subset \bigcup_{i=1}^{n} A_i$.

Lause 1.14. *Olgu S täielik meetriline ruum. Kui $K \subset S$ on täielikult tõkestatud, siis see on suhteliselt kompaktne.*

Tõestus. Olgu S täielik meetriline ruum ja K selle täielikult tõkestatud alamhulk. Olgu (x_n) mingi K elementide jada. Leidub arv $N \in \mathbb{N}$ ning hulgad $A_1, \dots, A_N \subset S$ nii, et $\operatorname{diam}(A_i) < 1$, $i = 1, \dots, m$, ja $K \subset \bigcup_{i=1}^{m} A_i$. Vähemalt üks hulk A_j peab sisaldama loenduva hulga jada (x_n) elemente. Tähistame $B_1 = A_j \cap K$ ja valime sellest hulgast konstrueeritava osajada esimese elemendi $x_{n_1} \in B_1$.
Kuna B_1 on täielikult tõkestatud hulga alamahulk, siis ka tema ise on täielikult tõkestatud, s.t leiduvad hulgad $C_1, \dots, C_l \subset S$ nii, et nende diameeter ei ületa $\frac{1}{2}$, kusjuures $B_1 \subset \bigcup_{i=1}^{l} C_i$. Taas peab leiduma vähemalt üks hulk C_k, et $C_k \cap B_1$ sisaldaks loenduva hulga jada (x_n) elemente. Tähistame $B_2 = C_k \cap B_1$ ning valime osajada teise elemendi $x_{n_2} \in B_2$ nii, et $n_2 > n_1$.
Jätkame sama protsessi. Sammul j vaatleme hulga B_{j-1} niisugust lõplikku katet $\bigcup_{i=1}^{n} D_i$, ei ühegi $i = 1, \dots, n$ korral hulga D_i diameeter ei ületa $\frac{1}{j}$. Saame (x_n) osajada, mis on Cauchy jada. Tõepoolest, fikseerime vabalt $\epsilon > 0$. Valime $N \in \mathbb{N}$ nii, et $\frac{1}{N} < \epsilon$. Kui $k, l \geq N$, siis $d(x_{n_k}, x_{n_l}) \leq \frac{1}{k} < \epsilon$. S täielikkuse tõttu osajada (x_{n_j}) koondub.
Kokkuvõttes oleme näidanud, et suvalisest hulga K elementide jadast saab eraldada koonduva osajada. Järelikult K on suhteliselt kompaktne. □

Teoreem 1.15. *Kui S on separaabel ja täielik, siis $\mathcal{P}$ on täielik.*

Tõestus. Tahame näidata, et $\mathcal{P}$ on täielik. Olgu $\{p_n\}$ mingi tõenäosusmõõtude Cauchy jada. Näitame, et $\{p_n\}$ on tihe. Siis Prohhorovi teoreemi põhjal on jada elementide hulk ka suhteliselt kompaktne. Järelikult sellest saab eraldada koonduva osajada ning seega $\{p_n\}$ koondub.

Fikseerime vabalt $\epsilon > 0$ ja $\delta > 0$. Valime $\eta > 0$ nii, et $2\eta < \min\{\epsilon, \delta\}$. Nüüd valime $N \in \mathbb{N}$ nii, et $n \geq N \Rightarrow \rho(p_N, p_n) < \eta$ (see on võimalik, sest tegemist on Cauchy jadaga). Kuna S on separaabel, leidub selles kõikjal tihe ülimalt loenduv alamhulk X. Paneme tähele, et $S = \bigcup_{x_i \in X} B(x_i, \eta)$. Tähistame $A_i = B(x_i, \eta)$. Olgu $k, l \in \mathbb{N}$. Siis $\bigcup_{i=1}^{k} A_i \subset \bigcup_{i=1}^{l} A_i$, kui $k \leq l$. Märgime, et $\bigcup_{i=1}^{\infty} A_i$ on mõõtuv, sest tegemist on lahtiste hulkade loenduva ühendiga. Niisiis, iga $n \in \mathbb{N}$ korral

$$\lim_k p_n\left(\bigcup_{i=1}^{k} A_i\right) = p_n\left(\bigcup_{i=1}^{\infty} A_i\right) = p_n(S) = 1.$$

See tähendab, et suvalise n puhul leidub m_n nii, et kehtib implikatsioon

$$k \geq m_n \Rightarrow 1 - p_n(\bigcup_{i=1}^{k} A_i) < \eta.$$

Tähistame $m := \max\{m_1, \ldots, m_N\}$. Siis $1 - p_n(\bigcup_{i=1}^{m} A_i) < \eta$ ehk $p_n(\bigcup_{i=1}^{m} A_i) > 1 - \eta$ iga $n \leq N$ korral.

Vaatleme lahtiseid kerasid $B_i = B(x_i, 2\eta)$ ning $C_i = B(x_i, \delta)$, kus $x_i \in X$, $i \in \mathbb{N}$. Kui $n \geq N$, siis

$$p_n\big(\bigcup_{i \leq m} C_i\big) \geq p_n\big(\bigcup_{i \leq m} B_i\big) \geq p_n\Big(\big(\bigcup_{i \leq m} A_i\big)^{\eta}\Big) \geq p_N\big(\bigcup_{i \leq m} A_i\big) - \eta \geq 1 - 2\eta > 1 - \epsilon.$$

Kui $n \leq N$, siis

$$p_n\big(\bigcup_{i \leq m} C_i\big) \geq p_n\big(\bigcup_{i \leq m} B_i\big) \geq p_n\big(\bigcup_{i \leq m} A_i\big) \geq 1 - \eta > 1 - \epsilon.$$

Lõpuks fikseerime vabalt $\xi > 0$. Suvalise $k \in \mathbb{N}$ korral olgu $\epsilon_k = \frac{\xi}{2^k}$ ning $\delta_k = \frac{1}{k}$. Moodustame kerad $C_{ki} = B(x_i, \delta_k)$, $i \in \mathbb{N}$, ning märkame, et need katavad ruumi S. Eelneva arutelu põhjal leidub m_k nii, et $p_n\big(\bigcup_{i \leq m_k} C_{ki}\big) > 1 - \frac{\xi}{2^k}$ iga $n \in \mathbb{N}$ korral.

Vaatleme hulka $K = \bigcap_{k \in \mathbb{N}} \bigcup_{i \leq m_k} C_{ki}$. Olgu $E > 0$ suvaline. Valime k nii, et $\delta_k < \frac{E}{2}$. Siis $\operatorname{diam}(C_{ki}) < E$, seejuures $K \subset \bigcup_{i \leq m_k} C_{ki}$. Niisiis, K on täielikult tõkestatud. Et S on täielik, siis K on teoreemi 1.14 põhjal suhteliselt kompaktne. Järelikult selle sulund $\overline{K}$ on kompaktne, kusjuures iga $n \in \mathbb{N}$ korral

$$p_n(\overline{K}) \geq p_n(K) = p_n(\bigcap_{k\in\mathbb{N}} \bigcup_{i\leq m_k} C_{ki}) = 1 - p_n(\bigcup_{k\in\mathbb{N}} \bigcap_{i\leq m_k} C_{ki}{}^{\mathsf{c}}) \geq$$
$$\geq 1 - \sum_{k\in\mathbb{N}} p_n(\bigcap_{i\leq m_k} C_{ki}{}^{\mathsf{c}}) = 1 - \sum_{k\in\mathbb{N}} p_n\Bigg(\Big(\bigcup_{i\leq m_k} C_{ki}\Big)^{\mathsf{c}}\Bigg) =$$
$$= 1 - \sum_{k\in\mathbb{N}} \big(1 - p_n(\bigcup_{i\leq m_k} C_{ki})\big) > 1 - \sum_{k\in\mathbb{N}} \frac{\xi}{2^k} = 1 - \xi$$

Järelikult $\{p_n\}$ on tihe ja ruum $\mathcal{P}$ on täielik. □

Järgnevalt käsitleme kujutusi, mis osutuvad väga olulisteks töövahenditeks Dirichlet' protsessi defineerimisel. Iga $A \in \mathcal{B}(S)$ korral defineerime kujutuse

$$T_A : \mathcal{P} \to [0,1], \quad p \mapsto p(A).$$

Tähistagu $C_b(S)$ kõigi pidevate tõkestatud funktsioonide $f : S \to \mathbb{R}$ hulka. Iga $f \in C_b(S)$ korral defineerime kujutuse

$$T_f : \mathcal{P} \to \mathbb{R}, \quad p \mapsto \int_S f dp.$$

Kujutuste T_A ja T_f üks väga tähtis omadus on see, et kõik need on mõõtuvad Boreli σ-algebra $\mathcal{B}(\mathcal{P}) = \sigma(\tau_\rho)$ suhtes. Järgnevate tulemuste tõestused tuginevad [5, lk 509–510] esitatud skeemidele.

Lause 1.16. ***(a)*** *Iga $f \in C_b(S)$ korral on funktsioon T_f $\mathcal{B}(\mathcal{P})$-mõõtuv ehk $\sigma(T_f; f \in C_b(S)) \subset \mathcal{B}(\mathcal{P})$.*

(b) *Iga $A \in \mathcal{B}(S)$ korral on funktsioon T_A $\mathcal{B}(\mathcal{P})$-mõõtuv ehk $\sigma(T_A; A \in \mathcal{B}(S)) \subset \mathcal{B}(\mathcal{P})$.*

Tõestus. **a)** Olgu $f \in C_b(S)$ suvaline. Fikseerime vabalt $p \in \mathcal{P}$. Olgu (p_n) mingi ruumi $\mathcal{P}$ kuuluvate elementide jada, mis koondub mõõduks p, s.t $\rho(p_n, p) \xrightarrow{n} 0$. Lause 1.7 põhjal järeldub Prohhorovi meetrika järgi koondumisest tõenäosusmõõtude nõrk koondumine, seega

$$T_f(p_n) = \int_S f dp_n \xrightarrow{n} \int_S f dp = T_f(p).$$

Niisiis pidevuse Heine kriteeriumi kohaselt on funktsioon T_f pidev.

Tähistame ruumi $\mathbb{R}$ kõigi lahtiste hulkade kogumit sümboliga $\tau_\mathbb{R}$. Et T_f on pidev, siis lahtiste hulkade originaalid on lahtised. Järelikult

$$\mathcal{F} := \{T_f^{-1}(G) : G \in \tau_\mathbb{R}\} \subset \tau_\rho.$$

Teame, et $\sigma(\mathcal{F}) = \{T_f^{-1}(G) : G \in \sigma(\tau_{\mathbb{R}}) = \mathcal{B}(\mathbb{R})\}$, seega

$$\{T_f^{-1}(G) : G \in \mathcal{B}(\mathbb{R})\} = \sigma(\mathcal{F}) \subset \sigma(\tau_\rho) = \mathcal{B}(\mathcal{P}).$$

Kuna suvalise $G \in \mathcal{B}(\mathbb{R})$ korral $T_f^{-1}(G) \in \mathcal{B}(\mathcal{P})$, siis T_f on $\mathcal{B}(\mathcal{P})$-mõõtuv.

b) Olgu $A \in \mathcal{B}(S)$ kinnine. Fikseerime vabalt $r \in \mathbb{R}$. Tahame näidata, et hulk

$$G \;:=\; \{p \in \mathcal{P} : T_A(p) < r\}$$

on lahtine. Selleks veendume, et $G^{\mathrm{c}} = \{p \in \mathcal{P} : T_A(p) \geq r\}$ on kinnine. Olgu (p_n) hulga G^{c} elementide mingi jada, mis koondub mõõduks p^*, s.t $\rho(p_n, p^*) \xrightarrow{n} 0$. Lause 1.7 põhjal leiab aset ka nõrk koondumine $p_n \Rightarrow p^*$. Et A on kinnine, siis Portmanteau' teoreemi kohaselt

$$r \leq \limsup_n T_A(p_n) = \limsup_n p_n(A) \leq p^*(A) = T_A(p^*).$$

Näeme, et $p^* \in G^{\mathrm{c}}$, seega hulk G^{c} on kinnine ja hulk G lahtine. Kuna

$$G = \{p \in \mathcal{P} : T_A(p) < r\} \in \tau_\rho \subset \mathcal{B}(\mathcal{P})$$

suvalise reaalarvu r korral, siis T_A on mõõtuv.

Nüüd eeldame, et $A \in \mathcal{B}(S)$ on lahtine. Veendume, et suvalise $r \in \mathbb{R}$ korral hulk

$$U := \{p \in \mathcal{P} : T_A(p) > r\}$$

on lahtine. Selleks näitame, et $U^{\mathrm{c}} = \{p \in \mathcal{P} : T_A(p) \leq r\}$ on kinnine. Olgu (p_n) hulga U^{c} elementide jada, mis koondub mõõduks p', s.t $\rho(p_n, p') \xrightarrow{n} 0$. Lause 1.7 põhjal leiab aset ka nõrk koondumine $p_n \Rightarrow p'$. Et A on lahtine, siis Portmanteau' teoreemi kohaselt

$$r \geq \liminf_n p_n(A) \geq p'(A) = T_A(p').$$

Niisiis, $p' \in U^{\mathrm{c}}$, mis tähendab, et U^{c} kinnine ja U lahtine. Et $U = \{p \in \mathcal{P} : T_A(p) > r\} \in \mathcal{B}(\mathcal{P})$ iga reaalarvu r korral, siis T_A on mõõtuv.

Veendume, et $\Lambda := \{A \in \mathcal{B}(S) : T_A \in m\mathcal{B}(\mathcal{P})\}$ on λ-süsteem.

1. S on kinnislahtine, seega $T_S \in m\mathcal{B}(\mathcal{P})$ ja $S \in \Lambda$.

2. Fikseerime vabalt $A \in \Lambda$. Tahame näidata, et ka $A^{\mathrm{c}} \in \Lambda$. Olgu $E \in \mathcal{B}(\mathbb{R})$ suvaline. Teame, et iga Boreli hulga nihe ja kordne on samuti Boreli hulk, seega $-1 \cdot (E - 1) \in \mathcal{B}(\mathbb{R})$. Järelikult

$$T_A^{-1}[-1 \cdot (E - 1)] = \{p \in \mathcal{P} : T_A(p) = p(A) \in -1 \cdot (E - 1)\} \in \mathcal{B}(\mathcal{P}).$$

Paneme tähele, et $p(A) \in -1 \cdot (E-1) \Leftrightarrow -p(A) \in E-1 \Leftrightarrow 1-p(A) \in E \Leftrightarrow p(A^c) \in E$. Niisiis,

$$\{p \in \mathcal{P} : T_{A^c}(p) \in E\} = \{p \in \mathcal{P} : T_A(p) \in -1 \cdot (E-1)\} \in \mathcal{B}(\mathcal{P}).$$

Oleme näidanud, et suvalise $E \in \mathcal{B}(\mathbb{R})$ korral $T_{A^c}^{-1}[E] \in \mathcal{B}(\mathcal{P})$, seega $A^c \in \Lambda$.

3. Olgu $A_i \in \Lambda, i \in \mathbb{N}$ sellised, et $A_i \cap A_j = \varnothing$, kui $i \neq j$. Tahame näidata, et ka $\bigcup_{i=1}^{\infty} A_i \in \Lambda$. Selleks peame veenduma, et $T_{\bigcup_{i=1}^{\infty} A_i}$ on mõõtuv.
Kõigepealt paneme tähele, et suvalise $p \in \mathcal{P}$ korral

$$T_{\bigcup_{i=1}^{\infty} A_i}(p) = p\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i\in\mathbb{N}} p(A_i) = \sum_{i\in\mathbb{N}} T_{A_i}(p).$$

Vaatleme funkstioone $g_n : \mathcal{P} \to \mathbb{R}$, $g_n := \sum_{i=1}^{n} T_{A_i}$, $n \in \mathbb{N}$. Iga $n \in \mathbb{N}$ korral g_n on mõõtuvate funktsioonide lõplik summa, järelikult ka ise mõõtuv. Märkame, et

$$\lim_{n\to\infty} g_n = \lim_{n\to\infty} \sum_{i=1}^{n} T_{A_i} = \sum_{i=1}^{\infty} T_{A_i} = T_{\bigcup_{i=1}^{\infty} A_i}.$$

Kuna $T_{\bigcup_{i=1}^{\infty} A_i}$ on mõõtuvate funktsioonide jada piirväärtus, siis on see ise mõõtuv funktsioon. Järelikult $\bigcup_{i=1}^{\infty} A_i \in \Lambda$.
Kuna oleme juba näidanud, et iga lahtise hulga $A \in \mathcal{B}(S)$ korral T_A on mõõtuv, siis on ilmne, et π-süsteem $\{A \in \mathcal{B}(S) : A \text{ on lahtine}\}$ sisaldub kogumis Λ. Dynkini teoreemi põhjal $\mathcal{B}(S) = \sigma(\{A \in \mathcal{B}(S) : A \text{ on lahtine}\}) \subset \Lambda$. See tähendab, et iga $A \in \mathcal{B}(S)$ korral T_A on $\mathcal{B}(\mathcal{P})$-mõõtuv. □

Äsja veendusime, et kõik kujutused T_f ja T_A on $\mathcal{B}(\mathcal{P})$-mõõtuvad. Juhul, kui S on Poola ruum, kehtib ka tugevam väide: $\mathcal{B}(\mathcal{P})$ on ühtlasi *vähim* σ-algebra, mille suhtes kõik kujutused T_f ja T_A on mõõtuvad ehk $\mathcal{B}(\mathcal{P}) = \sigma(T_A; A \in \mathcal{B}(S))$ ja $\mathcal{B}(\mathcal{P}) = \sigma(T_f; f \in C_b(S))$. Tõestuse hõlbustamiseks toome sisse nõrga topoloogia mõiste.

Definitsioon 1.17. Nõrk topoloogia. Nõrgimat topoloogiat ruumil $\mathcal{P}$, mille suhtes iga $f \in C_b(S)$ korral kujutus T_f on pidev, nimetatakse nõrgaks topoloogiaks. Tähitsame seda tähisega τ_w.

Lause 1.18. *Kui S on Poola ruum, siis kehtivad järgmised sisalduvused:*

(a) $\mathcal{B}(\mathcal{P}) \subset \sigma(T_f; f \in C_b(S))$;

(b) $\mathcal{B}(\mathcal{P}) \subset \sigma(T_A; A \in \mathcal{B}(S))$.

Tõestus. **a)** Paneme tähele, et kogum

$$\mathfrak{B} = \left\{ \bigcap_{i=1}^{n} T_{f_i}^{-1}((-\infty, c_i)) : n \in \mathbb{N}, f_i \in C_b(S), c_i \in \mathbb{R} \right\}$$

on nõrga topoloogia baas. Tõepoolest, valides suvalise $p \in \mathcal{P}$ korral mingi $f \in C_b(S)$, tähistame $r := \int_S f dp = T_f(p)$. Sellisel juhul $p \in T_f^{-1}((-\infty, r+1)) \in \mathfrak{B}$. On ilmne, et suvaliste $B_1, B_2 \in \mathfrak{B}$ korral ka $B_1 \cap B_2 \in \mathfrak{B}$. Järelikult $\mathfrak{B}$ on ruumi $\mathcal{P}$ mingi topoloogia baas. Kui τ on ruumi $\mathcal{P}$ topoloogia, mille suhtes iga T_f on pidev, siis suvalise $f \in C_b(S)$ ja $c \in \mathbb{R}$ korral $T_f^{-1}((-\infty, c)) \in \tau$, kuna $(-\infty, c)$ on lahtine $\mathbb{R}$ loomulikus topoloogias. Järelikult $\mathfrak{B} \subset \tau$, mis tähendab, et $\mathfrak{B}$ on nõrga topoloogia baas.
Märgime, et ρ indutseerib nõrga topoloogia ehk $\tau_\rho = \tau_w$. See fakt on tõestatud näiteks [2, lk 104–105]. Et S on eelduse kohaselt Poola ruum, siis teoreemi 1.9 põhjal on $(\mathcal{P}, \tau_\rho)$ separaabel. Selle tõttu on iga $\mathcal{P}$ τ_ρ-lahtine alamhulk esitatav kogumi $\mathfrak{B}$ hulkade loenduva ühendina. Järelikult $\tau_\rho \subset \sigma(\mathfrak{B})$. Kuna ilmselt $\mathfrak{B} \subset \sigma(T_f; f \in C_b(S))$, siis ka

$$\mathcal{B}(\mathcal{P}) = \sigma(\tau_\rho) = \sigma(\mathfrak{B}) \subset \sigma(T_f; f \in C_b(S)).$$

b) Nüüd vaatleme suvalist pidevat tõkestatud funktsiooni $f : S \to \mathbb{R}$. Eeldame, et see on mittenegatiivne (negatiivsete väärtustega funktsiooni saame vaadelda tema positiivse ja negatiivse osa vahena). Leiduvad mittenegatiivsed lihtsad mõõtuvad funktsioonid standardesitusega $\phi_n = \sum_{j=1}^{k_n} \alpha_j \chi_{A_j}$ nii, et $\phi_n \uparrow f$. Paneme tähele, et suvalise $p \in \mathcal{P}$ korral

$$T_{\phi_n}(p) = \int_S \phi_n dp = \sum_{j=1}^{k_n} \alpha_j p(A_j) = \sum_{i=1}^{k_n} \alpha_j T_{A_j}(p).$$

Et mõõtuvate funktsioonide lõplik summa on mõõtuv, siis iga $n \in \mathbb{N}$ korral funktsioon T_{ϕ_n} on $\sigma(T_A; A \in \mathcal{B}(S))$-mõõtuv. Järelikult iga $p \in \mathcal{P}$ puhul

$$T_f(p) = \int_S f dp = \int_S \lim_n \phi_n dp = \lim_n \int_S \phi_n dp = \lim_n T_{\phi_n}(p),$$

s.t ka T_f on $\sigma(T_A; A \in \mathcal{B}(S))$-mõõtuv. Näeme, et $\sigma(T_f; f \in C_b(S)) \subset \sigma(T_A; A \in \mathcal{B}(S))$. Järelikult osa a põhjal $\mathcal{B}(\mathcal{P}) \subset \sigma(T_A; A \in \mathcal{B}(S))$. □

Nüüdseks on meil olemas Boreli σ-algebra $\mathcal{B}(\mathcal{P})$, kusjuures oleme selgitanud välja ka olulise seose kujutustega T_A ja T_f. See võimaldab käsitleda tõenäosusmõõte ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$. Järgnevad teoreemid seovad tõenäosusmõõdud ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ mainitud kujutustega

T_A ja T_f. Esimese teoreemi intuitiivne sisu seisneb selles, et juhul, kui iga mõõtuva tükelduse $A_1, \ldots, A_k$ korral on teada juhusliku vektori $(T_{A_1}, \ldots, T_{A_k})$ jaotus, siis on määratud ka mõõt Π. Analoogiliselt, kui iga $f \in C_b(S)$ korral on teada juhusliku suuruse T_f jaotus, siis on määratud ka mõõt Π. Antud tulemusi rakendame Dirichlet' protsessi jaotuse käsitlemisel.

Teoreem 1.19. *Kui Π ja Π' on kaks erinevat tõenäosusmõõtu ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, siis leidub ruumi S mõõtuv tükeldus $A_1, \ldots, A_k$ ja k-mõõtmeline Boreli hulk $B^k \in \mathcal{B}(\mathbb{R}^k)$ nii, et*

$$\Pi\{p : (T_{A_1}(p), \ldots, T_{A_k}(p)) \in B^k\} \neq \Pi'\{p : (T_{A_1}(p), \ldots, T_{A_k}(p)) \in B^k\}.$$

Tõestus. Eeldame, et iga hulga S mõõtuva tükelduse $A_1, \ldots, A_k$ ja k-mõõtmelise Boreli hulga $B^k \in \mathcal{B}(\mathbb{R}^k)$ korral

$$\Pi\{p : (T_{A_1}(p), \ldots, T_{A_k}(p)) \in B^k\} = \Pi'\{p : (T_{A_1}(p), \ldots, T_{A_k}(p)) \in B^k\}.$$

Tahame näidata, et mõõdud Π ja Π' on võrdsed.
Hulga S iga mõõtuva tükelduse $A_1, \ldots, A_k$ puhul defineerime σ-algebra ruumil $\mathcal{P}$:

$$\mathcal{F}_{A_1,\ldots,A_k} = \sigma(T_{A_1}, \ldots, T_{A_k}).$$

Olgu $\mathcal{T}$ ruumi S kõikvõimalike lõplike mõõtuvate tükelduste hulk. Defineerime kogumi

$$\mathfrak{A} = \bigcup_{t \in \mathcal{T}} \mathcal{F}_t.$$

Veendume, et tegemist on algebraga.

A1. Selge, et $\varnothing, \mathcal{P} \in \mathfrak{A}$, sest $\varnothing, \mathcal{P} \in \mathcal{F}_t$ iga $t \in \mathcal{T}$ korral.

A2. Olgu $A \in \mathfrak{A}$. Siis leidub hulga S mõõtuv tükeldus $t \in \mathcal{T}$ nii, et $A \in \mathcal{F}_t \subset \mathfrak{A}$. Järelikult ka $A^c \in \mathcal{F}_t \subset \mathfrak{A}$.

A3. Olgu $A, B \in \mathfrak{A}$. Leiduvad hulga S mõõtuvad tükeldused $A_1, \ldots, A_k$ ja $B_1, \ldots, B_m$, et $A \in \mathcal{F}_{A_1,\ldots,A_k}$ ja $B \in \mathcal{F}_{B_1,\ldots,B_m}$. Moodustame hulgad $C_{ij} = A_i \cap B_j$, kus $i \in \{1, \ldots, k\}$ ja $j \in \{1, \ldots, m\}$. Iga C_{ij} on mõõtuv, sest see on kahe mõõtuva hulga ühisosa. Selge, et hulgad C_{ij} on omavahel lõikumatud ja $\bigcup_{1 \leq j \leq m} \bigcup_{1 \leq i \leq k} C_{ij} = S$ ehk tegemist on hulga S mõõtuva tükeldusega, mida tähistame tähega t^*.

Paneme tähele, et $\mathcal{F}_{A_1,\dots,A_k} \subset \mathcal{F}_{t*}$ ja $\mathcal{F}_{B_1,\dots,B_m} \subset \mathcal{F}_{t*}$. Tõepoolest, suvalise $i \in \{1,\dots,k\}$ korral $A_i = \bigcup\limits_{j=1}^{m} C_{ij}$. Järelikult iga $p \in \mathcal{P}$ korral

$$T_{A_i}(p) = T_{\bigcup_{j=1}^m C_{ij}}(p) = p(\bigcup_{j=1}^{m} C_{ij}) = \sum_{j=1}^{m} p(C_{ij}) = \sum_{j=1}^{m} T_{C_{ij}}(p).$$

Et kõik funktsioonid $T_{C_{ij}}$ on $\mathcal{F}_{t^*}$ suhtes mõõtuvad, siis ka iga nende summa on mõõtuv. Niisiis, T_{A_i} on $\mathcal{F}_{t^*}$-mõõtuv. Sama mõttekäik kehtib ka funktsioonide T_{B_j}, $j \in \{1,\dots,m\}$, korral.

Kokkuvõttes $A, B \in \mathcal{F}_{t^*}$. Et $\mathcal{F}_{t^*}$ on σ-algebra, siis ka $A \cup B \in \mathcal{F}_{t^*} \subset \mathfrak{A}$.

Eelduse kohaselt kehtib iga hulga S mõõtuva tükelduse $A_1,\dots,A_k$ korral võrdus

$$\Pi'\big|_{\mathcal{F}_{A_1,\dots,A_k}} = \Pi\big|_{\mathcal{F}_{A_1,\dots,A_k}}.$$

Olgu $A \in \mathfrak{A}$ suvaline. Leidub tükeldus $t \in \mathcal{T}$ nii, et $A \in \mathcal{F}_t$. Järelikult

$$\Pi'(A) = \Pi'\big|_{\mathcal{F}_t}(A) = \Pi\big|_{\mathcal{F}_t}(A) = \Pi(A)$$

ja seega $\Pi\big|_{\mathfrak{A}} = \Pi'\big|_{\mathfrak{A}} =: \Pi_0$. Paneme tähele, et $\sigma(\mathfrak{A})$ on σ-algebra, mille suhtes kõik funktsioonid $T_A, A \in \mathcal{B}(S)$ on mõõtuvad. Lausete 1.16 ja 1.18 põhjal $\mathcal{B}(\mathcal{P}) = \sigma(\mathfrak{A})$.
Olgu Π_0^* mõõduga Π_0 assotsieeruv välismõõt. Hahni jätkamisteoreemi põhjal $\Pi_0^*\big|_{\sigma(\mathfrak{A})}$ on mõõdu Π_0 ainuke jätk σ-agebrale $\sigma(\mathfrak{A})$. See tähendab, et iga $A \in \mathcal{B}(\mathcal{P})$ korral

$$\Pi(A) = \Pi_0^*\big|_{\sigma(\mathfrak{A})}(A) = \Pi'(A).$$

□

Teoreemi 1.21 tõestamiseks toome sisse lause 1.20. Selle tõesuse leiab nt [6, lk 202].

Lause 1.20. *Juhuslikud vektorid $(X_1,\dots,X_n)$ ja $(Y_1,\dots,Y_n)$ on erineva jaotusega parajasti siis, kui leiduvad arvud $a_1,\dots,a_k$ nii, et juhuslikud suurused $\sum\limits_{i=1}^{k} a_i X_i$ ja $\sum\limits_{i=1}^{k} a_i Y_i$ on erineva jaotusega.*

Teoreem 1.21. *Kui Π ja Π' on kaks erinevat tõenäosusmõõtu ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, siis leidub funktsioon $f \in C_b(S)$ ja Boreli hulk $B \in \mathcal{B}(\mathbb{R})$ nii, et*

$$\Pi\{p : T_f \in B\} \neq \Pi'\{p : T_f \in B\}.$$

Tõestus. Iga $k \in \mathbb{N}$ ja funktsioonide komplekti $f_1, \dots, f_k \in C_b(S)$ korral defineerime σ-algebra ruumil $\mathcal{P}$:

$$\mathcal{F}_{f_1,\dots,f_k} = \sigma(T_{f_1}, \dots, T_{f_k}).$$

Eeldame, et iga $f \in C_b(S)$ korral $\Pi\big|_{\mathcal{F}_f} = \Pi'\big|_{\mathcal{F}_f}$.
Olgu $\mathcal{C}$ kõikvõimalike selliste lõplike funktsioonide komplektide hulk. Defineerime kogumi

$$\mathfrak{A} = \bigcup_{c \in \mathcal{C}} \mathcal{F}_c.$$

Veendume, et $\mathfrak{A}$ on algebra.

A1. Selge, et $\varnothing, \mathcal{P} \in \mathfrak{A}$, sest $\varnothing, \mathcal{P} \in \mathcal{F}_c$ iga $c \in \mathcal{C}$ korral.

A2. Olgu $A \in \mathfrak{A}$. Siis leidub funktsioonide komplekt $c \in \mathcal{C}$ nii, et $A \in \mathcal{F}_c$. Järelikult ka $A^c \in \mathcal{F}_c \subset \mathfrak{A}$.

A3. Olgu $A, B \in \mathfrak{A}$. Siis leiduvad funktsioonid $f_1, \dots, f_k, g_1, \dots, g_m \in C_b(S)$ nii, et $A \in \mathcal{F}_{f_1,\dots,f_k}$ ja $B \in \mathcal{F}_{g_1,\dots,g_m}$. Seejuures on ilmne, et $\mathcal{F}_{f_1,\dots,f_k} \subset \mathcal{F}_{f_1,\dots,f_k,g_1,\dots,g_m}$ ja $\mathcal{F}_{g_1,\dots,g_m} \subset \mathcal{F}_{f_1,\dots,f_k,g_1,\dots,g_m}$. Niisiis, $A, B \in \mathcal{F}_{f_1,\dots,f_k,g_1,\dots,g_m}$ ja kuna tegemist on σ-algebraga, siis ka $A \cup B \in \mathcal{F}_{f_1,\dots,f_k,g_1,\dots,g_m} \subset \mathfrak{A}$.

Eelduse kohaselt kehtib iga $f \in C_b(S)$ puhul $\Pi\big|_{\mathcal{F}_f} = \Pi'\big|_{\mathcal{F}_f}$. Oletame vastuväiteliselt, et mingi funktsioonide komplekti $f_1, \dots, f_k \in C_b(S)$ korral

$$\Pi\big|_{\mathcal{F}_{f_1,\dots,f_k}} \neq \Pi'\big|_{\mathcal{F}_{f_1,\dots,f_k}}.$$

See tähendab, et leidub hulk $A \in \mathcal{F}_{f_1,\dots,f_k}$, mille korral

$$\Pi(A) \neq \Pi'(A).$$

Kuna $\mathcal{F}_{f_1,\dots,f_k}$ on vähim σ-algebra, mille suhtes kujutused $T_{f_1}, \dots, T_{f_k}$ on mõõtuvad, siis see on ka vähim σ-algebra, mille suhtes juhuslik vektor $(T_{f_1}, \dots, T_{f_k})$ on mõõtuv. Järelikult leidub $B \in \mathcal{B}(\mathbb{R})$, et $A = (T_{f_1}, \dots, T_{f_k})^{-1}(B)$. Niisiis,

$$\Pi(A) = \Pi(T_{f_1}, \dots, T_{f_k})^{-1}(B) \neq \Pi'(T_{f_1}, \dots, T_{f_k})^{-1}(B) = \Pi'(A).$$

Näeme, et $\Pi(T_{f_1}, \dots, T_{f_k})^{-1}$ ja $\Pi'(T_{f_1}, \dots, T_{f_k})^{-1}$ on juhusliku vektori $(T_{f_1}, \dots, T_{f_k})$ erinevad jaotused. Lause 1.20 põhjal leiduvad arvud $a_1, \dots, a_k \in \mathbb{R}$ ja hulk $C \in \mathcal{B}(\mathbb{R})$ nii, et

$$\Pi\Big(\sum_{i=1}^{k} a_i T_{f_i}\Big)^{-1}(C) \neq \Pi'\Big(\sum_{i=1}^{k} a_i T_{f_i}\Big)^{-1}(C).$$

Nüüd märkame, et suvalise $p \in \mathcal{P}$ korral

$$\sum_{i=1}^{k} a_i T_{f_i}(p) = \sum_{i=1}^{k} a_i \int_S f_i dp = \int_S \sum_{i=1}^{k} a_i f_i dp = T_{\sum_{i=1}^{k} a_i f_i}(p).$$

Pidevate tõkestatud funktsioonide lineaarkombinatsioon on aga samuti pidev tõkestatud funktsioon. Tähistame $g = \sum_{i=1}^{k} a_i f_i$. Selge, et $T_g^{-1}(C) \in \mathcal{F}_g$, kusjuures

$$\Pi(T_g^{-1}(C)) \neq \Pi'(T_g^{-1}(C)).$$

Oleme näidanud, et $\Pi\big|_{\mathcal{F}_g} \neq \Pi'\big|_{\mathcal{F}_g}$. Ent see on vastuolus eeldusega. Järelikult iga funkstioonide komplekti $f_1, \dots, f_k \in C_b(S)$ korral $\Pi\big|_{\mathcal{F}_{f_1,\dots,f_k}} = \Pi'\big|_{\mathcal{F}_{f_1,\dots,f_k}}$.

Olgu $A \in \mathfrak{A}$ suvaline. Siis leidub funktsioonide komplekt $c \in \mathcal{C}$ nii, et $A \in \mathcal{F}_c$. Järelikult $\Pi(A) = \Pi\big|_{\mathcal{F}_c}(A) = \Pi'\big|_{\mathcal{F}_c}(A) = \Pi'(A)$ ja seega $\Pi\big|_{\mathfrak{A}} = \Pi'\big|_{\mathfrak{A}} =: \Pi_0$. Paneme tähele, et $\sigma(\mathfrak{A})$ on σ-algebra, mille suhtes kõik funktsioonid T_f, $f \in C_b(S)$, on mõõtuvad. Lausete 1.16 ja 1.18 põhjal $\mathcal{B}(\mathcal{P}) = \sigma(\mathfrak{A})$.

Olgu nüüd Π_0^* mõõduga Π_0 assotsieeruv välismõõt. Hahni jätkamisteoreemi põhjal $\Pi_0^*\big|_{\sigma(\mathfrak{A})}$ on mõõdu Π_0 ainuke jätk σ-algebrale $\sigma(\mathfrak{A})$. Järelikult iga $A \in \mathcal{B}(\mathcal{P})$ korral

$$\Pi(A) = \Pi_0^*\big|_{\sigma(\mathfrak{A})}(A) = \Pi'(A).$$

$\square$

2. Lõplikumõõtmeline Dirichlet' jaotus

Lõplikumõõtmeline Dirichlet' jaotus leiab laialdast rakendust Bayesi statistikas ning sellele tugineb ka Dirichlet' protsessi mõiste. Käesolevas alapeatükis esitame lõplikumõõtmelise Dirichlet' jaotuse definitsiooni ning anname ülevaate omadustest, mis teevad sellest kasuliku töövahendi Bayesi raamistikus. Lõpuks esitame lihtsa algoritmi k-mõõtmelise Dirichlet' jaotusega juhusliku vektori genereerimiseks. Antud peatüki materjal tugineb põhiliselt [5, lk 562–570].

2.1 Definitsioon ja tähtsamad omadused

Definitsioon 2.1. Dirichlet' jaotuseks parameetriga $k \in \mathbb{N} \setminus \{1\}$ ja $\alpha = (\alpha_1, \ldots, \alpha_k) > 0$ nimetatakse sellist juhusliku vektori $(\mathbf{p}_1, \ldots, \mathbf{p_k})$, $\sum\limits_{i=1}^{k} \mathbf{p_i} = 1$, jaotust, mille tihedusfunktsioon $\pi(p) = \pi(p_1, \ldots, p_k)$ avaldub kujul

$$\pi(p_1, \ldots, p_k) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}, \quad p = (p_1, \ldots, p_k) \in \mathbb{S}_k,$$

kus $\mathbb{S}_k = \{(p_1, \ldots, p_k) : \min\limits_i p_i \geq 0, \sum\limits_{i=1}^{k} p_i = 1\}$.

Märkus 2.2. Märgime, et nõude $\sum\limits_{i=1}^{k} \mathbf{p_i} = 1$ tõttu saame esitada juhusliku vektori $(\mathbf{p}_1, \ldots, \mathbf{p_k})$ mistahes koordinaadi teiste koordinatide kaudu: $\mathbf{p_j} = 1 - \sum\limits_{\substack{i=1, \\ i \neq j}}^{k} \mathbf{p_i}$. Niisiis, on ülal antud tihedus sisuliselt $(k-1)$-mõõtmeline funktsioon.

Märkus 2.3. Erijuhul $k = 2$ nimetatakse Dirichlet' jaotust beetajaotuseks. Parameetreid α_1 ja α_2 tähistatakse sageli hoopis tähtedega α ja β. Lihtsuse mõttes vaadeldakse beetajaotust lõigus $[0, 1]$ väärtusi omandava juhusliku suuruse $\mathbf{p}_1$ ja mitte juhusliku vektori $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2)$ jaotusena, sest teine koordinaat on alati esitatav esimese kaudu, nimelt

$\mathbf{p}_2 = 1 - \mathbf{p}_1$. Selle tõttu esitatakse ka beetajaotuse tihedus ühe muutuja funktsioonina kujul

$$\pi(p_1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_1^{\alpha-1}(1-p_1)^{\beta-1}, \quad p_1 \in [0,1].$$

Asjaolu, et juhuslik suurus $\mathbf{p}_1$ on beetajaotusega, tähistame $\mathbf{p}_1 \sim \mathrm{Be}(\alpha,\beta)$.

Üks väga tähtis lõplikumõõtmelise Dirichlet' jaotuse omadus, millele tugineb ka Dirichlet' protsessi konstruktsioon, on ühenduvus (*aggregation*). Sisuliselt tähendab ühenduvus seda, et juhul, kui $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p_k})$ on $\mathrm{Dir}(k;\alpha_1,\ldots,\alpha_k)$ jaotusega juhuslik vektor, siis koordinaatide liitmisel moodustatud $(k-1)$-mõõtmeline vektor $(\mathbf{p}_1, \ldots \mathbf{p_i} + \mathbf{p_j}, \ldots \mathbf{p_k})$ on jaotusega $\mathrm{Dir}(k-1;\alpha_1,\ldots \alpha_i+\alpha_j,\ldots \alpha_k)$. Järgmise teoreemi tõestus on esitatud näiteks [5, lk 563–564].

Teoreem 2.4. *Ühenduvus.* *Olgu* $\mathbf{p} = (\mathbf{p}_1,\ldots,\mathbf{p_k})$ *Dir*$(k;\alpha_1,\ldots,\alpha_k)$ *jaotusega juhuslik vektor ja* $I_1,\ldots,I_m$ *indeksite hulga* $\{1,\ldots,k\}$ *lõikumatu tükeldus. Iga* $j \in \{1,\ldots,m\}$ *korral olgu* $\mathbf{p_j^*} = \sum_{i\in I_j} \mathbf{p_i}$*. Siis juhuslik vektor* $(\mathbf{p_1^*},\ldots,\mathbf{p_m^*})$ *on jaotusega Dir*$(m;\beta_1,\ldots,\beta_m)$*, kus* $\beta_j = \sum_{i\in I_j} \alpha_i$ *iga* $j \in \{1,\ldots,m\}$ *korral.*

Järgnevalt uurime olulist Dirichlet' jaotuse omadust, mis teeb sellest kasuliku töövahendi Bayesi statistilises anlüüsis. Tugevalt üldistades saab Bayesi lähenemisviisi puhul eristada kahte etappi. Kõigepealt postuleeritakse juhusliku vektori $\mathbf{p}$ jaotus, mida nimetatakse vastavalt $\mathbf{p}$ eeljaotuseks. Vektori $\mathbf{p}$ fikseeritud väärtuse $p = (p_1,\ldots,p_k)$ korral on $X_1,\ldots,X_n$ juhuslikud suurused jaotusega p. Olemasolevate andmete $X_1 = i_1,\ldots,X_n = i_n$ valguses uuendatakse arusaama $\mathbf{p}$ jaotusest ehk leitakse tinglik jaotus $\mathbf{p}|X_1 = i_1,\ldots,X_n = i_n$, mida nimetatakse $\mathbf{p}$ järeljaotuseks. Osutub, et juhul, kui parameetri $\mathbf{p}$ eeljaotus on Dirichlet' jaotus, siis teatud tingimustel on ka $\mathbf{p}$ järeljaotus Dirichlet' jaotus. Viimane omadus teeb järeljaotuse leidmise matemaatiliselt väga lihtsaks, kuna selle üldine kuju on ette teada ja uuendada tuleb üksnes parameetreid $\alpha_1,\ldots,\alpha_k$. Omaduse mõtestamiseks defineerime kõigepealt Dirichlet'-kategoorilise protsessi.

Definitsioon 2.5. Dirichlet'-kategooriline protsess. Olgu n mingi naturaalarv ja $\alpha = (\alpha_1,\ldots,\alpha_k) \in \mathbb{R}^k$. Vaatleme juhuslikku vektorit $(\mathbf{p}, X_1,\ldots,X_n)$, kus $\mathbf{p}$ on omakorda juhuslik vektor jaotusega $\mathrm{Dir}(k;\alpha)$. Fikseeritud juhusliku vektori $\mathbf{p}$ väärtuse $p = (p_1,\ldots,p_k)$ korral on $X_1,\ldots,X_n$ sõltumatud juhuslikud suurused jaotusega p (tähistame i.i.d), s.t iga $i \in \{1,\dots,n\}$ puhul X_i on juhuslik suurus väärtuste hulgaga $\{1,\dots,k\}$, kusjuures iga $j \in \{1,\dots,k\}$ korral $P(X_i = j) = p_j$ ning suvaliste $i_1,\dots,i_n \in \{1,\dots,k\}$ korral $P(X_1 = i_1,\dots,X_n = i_n|\mathbf{p} = p) = p_{i_1}\dots p_{i_n}$. Niisugust mudelit nimetatakse Dirichlet'-kategooriliseks protsessiks ja tähistatakse järgmiselt:

$$\mathbf{p}|\alpha, k \sim \mathrm{Dir}(k;\alpha),$$
$$X_1,\dots,X_n|\mathbf{p} = p \overset{\text{i.i.d}}{\sim} p.$$

Definitsioonis 2.5 kirjeldatud protsessis on $\mathrm{Dir}(k;\alpha_1,\dots,\alpha_k)$ juhusliku vektori $\mathbf{p}$ eeljaotus. Intuitiivselt kujutab see endast meie arusaama vektori $\mathbf{p}$ jaotusest enne vaatluste $X_1,\dots,X_n$ tegemist. Ent huvi pakub ka see, milline on vektori $\mathbf{p}$ jaotus vaatluste $X_1,\dots,X_n$ tegemise järel, s.t vektori $\mathbf{p}$ tinglik jaotus $\mathbf{p} \mid X_1,\dots,X_n$ ehk $\mathbf{p}$ järeljaotus. Järgnev teoreem annab vastuse sellele küsimusele.

Teoreem 2.6. *Olgu* $\mathbf{p}|\alpha, k \sim Dir(k;\alpha)$ *ja* $X_1,\dots,X_n|\mathbf{p} = p \overset{i.i.d}{\sim} p$. *Siis juhusliku vektori* $\mathbf{p}$ *järeljaotus on* $Dir(k;\alpha_1 + N_1^n,\dots,\alpha_k + N_k^n)$ *ehk*

$$\mathbf{p} \mid X_1,\dots,X_n \sim Dir(k;\alpha_1 + N_1^n,\dots,\alpha_k + N_k^n),$$

kus $N_j^n = \sum_{i=1}^{n} I_j(X_i)$.

Tõestus. Olgu $\pi(p) = \pi(p_1,\dots,p_k)$ Dirichlet' jaotuse tihedusfunktsioon ehk eelmõõdu tihedus ja $\pi(p|X_1 = i_1,\dots,X_n = i_n)$ tingliku jaotuse ehk järelmõõdu tihedus. Siis Bayesi teoreemi põhjal

$$\pi(p|X_1 = i_1,\dots,X_n = i_n) = \frac{\pi(p)P(X_1 = i_1,\dots,X_n = i_n|\mathbf{p} = p)}{P(X_1 = i_1,\dots,X_n = i_n)}.$$

Nüüd leiame nimetajas oleva tõenäosuse.

$$P(X_1 = i_1,\dots,X_n = i_n) = \int_{\mathbb{S}_k} P(X_1 = i_1,\dots,X_n = i_n|\mathbf{p} = (p_1,\dots,p_k))\pi(p)dp =$$

$$= \int_{\mathbb{S}_k} p_{i_1}\dots p_{i_n}\pi(p)dp = \int_{\mathbb{S}_k} p_1^{N_1^n}\dots p_k^{N_k^n}\cdot\frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\cdot p_1^{\alpha_1-1}\dots p_k^{\alpha_k-1}dp =$$

$$= \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\int_{\mathbb{S}_k} p_1^{N_1^n}\dots p_k^{N_k^n}\cdot p_1^{\alpha_1-1}\dots p_k^{\alpha_k-1}dp =$$

$$= \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\int_{\mathbb{S}_k} p_1^{\alpha_1+N_1^n-1}\dots p_{k-1}^{\alpha_k+N_k^n-1}dp =$$

$$= \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\cdot\frac{\Gamma(\alpha_1+N_1^n)\ldots\Gamma(\alpha_k+N_k^n)}{\Gamma((\alpha_1+N_1^n)+\ldots+(\alpha_k+N_k^n))}.$$

Märgime, et viimane võrdus kehtib, sest võrduse vasakul pool olev integraal on parajasti normaliseerimiskonstandi pöördarvu Dirichlet' vorm. Naaseme esialgse Bayesi teoreemist saadud võrduse juurde ja rakendame viimast tulemust järelmõõdu tiheduse leidmisel.

$$\pi(p|X_1=i_1,\ldots,X_n=i_n)=\frac{\pi(p)P(X_1=i_1,\ldots,X_n=i_n|\mathbf{p}=p)}{P(X_1=i_1,\ldots,X_n=i_n)}=$$

$$=\frac{\frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\prod_{i=1}^k p_i^{\alpha_i-1}\cdot p_{i_1}\cdot\ldots\cdot p_{i_n}}{\frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\cdot\frac{\Gamma(\alpha_1+N_1^n)\ldots\Gamma(\alpha_k+N_k^n)}{\Gamma((\alpha_1+N_1^n)+\ldots+(\alpha_k+N_k^n))}}=$$

$$=\frac{\prod_{i=1}^k p_i^{\alpha_i-1}\cdot p_1^{N_1^n}\cdot\ldots\cdot p_k^{N_k^n}}{\frac{\Gamma(\alpha_1+N_1^n)\ldots\Gamma(\alpha_k+N_k^n)}{\Gamma((\alpha_1+N_1^n)+\ldots+(\alpha_k+N_k^n))}}=$$

$$=\frac{\Gamma((\alpha_1+N_1^n)+\ldots+(\alpha_k+N_k^n))}{\Gamma(\alpha_1+N_1^n)\ldots\Gamma(\alpha_k+N_k^n)}\prod_{i=1}^k p_i^{\alpha_i+N_i^n-1}.$$

Näeme, et tegemist on parajasti $\mathrm{Dir}(k;\alpha_1+N_1^n,\ldots,\alpha_k+N_k^n)$ tihedusega. □

Järgnevalt veendume, et juhul, kui $X_1,\ldots,X_n$ on vaatlused Dirichlet' kategoorilisest protsessist, ei sõltu $\mathbf{p}$ järeljaotus nende vaatluste järjekorrast, vaid üksnes saavutatud väärtustest. Selleks toome sisse multinoomjaotuse mõiste.

Definitsioon 2.7. Multinoomjaotus. Olgu $X_1,\ldots,X_n$ sõltumatud sama jaotusega juhuslikud suurused väärtuste hulgaga $\{1,\ldots,k\}$, kusjuures iga $i\in\{1,\ldots,n\}$ korral $P(X_i=j)=p_j$ ja $\sum_{j=1}^k p_j=1$. Iga $j\in\{1,\ldots,k\}$ korral tähistagu R_j seerias $X_1,\ldots,X_n$ väärtuse j saavutanud vaatluste arvu. Multinoomjaotuseks parameetritega $n\in\mathbb{N}$ ja $p=(p_1,\ldots,p_k)$ nimetatakse juhusliku vektori $R=(R_1,\ldots,R_k)$ jaotust. Multinoomjaotuse tõenäosusfunktsioon $t(r)=t(r_1,\ldots,r_n)$ avaldub kujul

$$t(r_1,\ldots,r_k)=\frac{n!}{r_1!\ldots r_k!}p_1^{r_1}\ldots p_k^{r_k},$$

kus $r_1\ldots,r_k\in\mathbb{N}\cup\{0\}$ ja $\sum_{i=1}^k r_i=n$. Asjaolu, et juhuslik vektor $R=(R_1,\ldots,R_k)$ on multinoomjaotusega parameetritega n ja $p=(p_1,\ldots,p_k)$, tähistame $R\sim M(n;p)$. Multinoomjaotusega juhusliku vektori keskväärtus on $E(R)=(np_1,\ldots,np_k)$.

Järeldus 2.8. Olgu $\mathbf{p} \sim \mathrm{Dir}(k;\alpha_1,\ldots,\alpha_k)$ ja juhusliku vektori $\mathbf{p}$ fikseeritud väärtuse $p = (p_1,\ldots,p_k)$ korral $R = (R_1,\ldots,R_k) \sim M(n;p)$. Kui $R_1 = r_1,\ldots,R_k = r_k$, siis vektori $\mathbf{p}$ järeljaotus on $\mathrm{Dir}(k;\alpha_1 + r_1,\ldots,\alpha_k + r_k)$ ehk

$$\mathbf{p} \mid R_1 = r_1,\ldots,R_k = r_k \sim \mathrm{Dir}(k;\alpha_1 + r_1,\ldots,\alpha_k + r_k).$$

Tõestus. Bayesi teoreemi kohaselt

$$\pi(p|R_1 = r_1,\ldots,R_k = r_k) = \frac{P(R_1 = r_1,\ldots,R_k = r_k|\mathbf{p} = p)\pi(p)}{P(R_1 = r_1,\ldots,R_k = r_k)} =$$

$$= \frac{\frac{n!}{r_1!\ldots r_k!}p_1^{r_1}\ldots p_k^{r_k}\pi(p)}{\int_{\mathbb{S}_k} P(R_1 = r_1,\ldots,R_k = r_k|\mathbf{p} = p)\pi(p)dp} = \frac{\frac{n!}{r_1!\ldots r_k!}p_1^{r_1}\ldots p_k^{r_k}\pi(p)}{\frac{n!}{r_1!\ldots r_k!}\int_{\mathbb{S}_k} p_1^{r_1}\ldots p_k^{r_k}\pi(p)dp} =$$

$$= \frac{p_1^{r_1}\ldots p_k^{r_k}\pi(p)}{\int_{\mathbb{S}_k} p_1^{r_1}\ldots p_k^{r_k}\pi(p)dp}.$$

Olgu nüüd $X_1,\ldots,X_n$ sõltumatud juhuslikud vaatlused jaotusest p, mis määravad vektori $(R_1,\ldots,R_k)$, s.t $R_j = \sum_{i=1}^{n} I_j(X_i)$ iga $j \in \{1,\ldots,k\}$ korral. Märkame, et

$$\frac{p_1^{r_1}\ldots p_k^{r_k}\pi(p)}{\int_{\mathbb{S}_k} p_1^{r_1}\ldots p_k^{r_k}\pi(p)dp} = \frac{P(X_1 = i_1,\ldots,X_n = i_n|\mathbf{p} = p)\pi(p)}{\int_{\mathbb{S}_k} P(X_1 = i_1,\ldots,X_n = i_n|\mathbf{p} = p)\pi(p)dp} =$$

$$= \frac{P(X_1 = i_1,\ldots,X_n = i_n|\mathbf{p} = p)\pi(p)}{P(X_1 = i_1,\ldots,X_n = i_n)} = \pi(p|X_1 = i_1,\ldots,X_n = i_n).$$

Viimane on aga juhusliku vektori $\mathbf{p}$ järeljaotuse tihedusfunktsioon Dirichlet'-kategoorilises protsessis. Niisiis, teoreemi 2.6 kohaselt on see parajasti jaotuse $\mathrm{Dir}(k;\alpha_1+r_1,\ldots,\alpha_k+r_k)$ tihedusfunktsioon. □

Nüüd võtame vaatluse alla Dirichlet' jaotusega juhusliku vektori keskväärtuse ja selle koordinaatide dispersiooni. Kõigepealt tõestame laused nende üldise kuju kohta ning seejärel anname põgusa ülevaate kaasnevatest eelistest Bayesi statistilise analüüsi raamistikus.

Lause 2.9. ***Keskväärtus.*** *Olgu* $\mathbf{p} = (\mathbf{p}_1,\ldots,\mathbf{p_k})$ *juhuslik vektor jaotusega* $Dir(k;\alpha_1,\ldots,\alpha_k)$. *Siis vektori* $\mathbf{p}$ *keskväärtus on*

$$E(\mathbf{p}) = \left(\frac{\alpha_1}{|\alpha|},\ldots,\frac{\alpha_k}{|\alpha|}\right),$$

kus $|\alpha| = \sum_{i=1}^{k}\alpha_i$.

Tõestus. Olgu $j \in \{1, \dots, k\}$ suvaline. Keskväärtuse definitsiooni kohaselt

$$E(\mathbf{p_j}) = \int_{\mathbb{S}_k} p_j \cdot \pi(p) dp = \int_{\mathbb{S}_k} p_j \cdot \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \prod_{i=1}^{k} p_i^{\alpha_i - 1} dp =$$

$$= \int_{\mathbb{S}_k} \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \prod_{\substack{i=1,\\ i \neq j}}^{k-1} p_i^{\alpha_i - 1} \cdot p_j^{\alpha_j} dp.$$

Gammafunktsiooni omaduse kohaselt $\Gamma(\alpha_1 + \dots + \alpha_k + 1) = |\alpha| \cdot \Gamma(\alpha_1 + \dots + \alpha_k)$, samuti $\Gamma(\alpha_j + 1) = \alpha_j \Gamma(\alpha_j)$. Tähistame $\beta_j = \alpha_j + 1$. Niisiis, saame kirjutada

$$E(\mathbf{p_j}) = \int_{\mathbb{S}_k} \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \prod_{\substack{i=1,\\ i \neq j}}^{k-1} p_i^{\alpha_i - 1} \cdot p_j^{\alpha_j} dp =$$

$$= \int_{\mathbb{S}_k} \frac{\alpha_j}{|\alpha|} \cdot \frac{\Gamma(\alpha_1 + \dots + \beta_j + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\beta_j) \dots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \dots p_j^{\beta_j - 1} \dots p_j^{\alpha_k - 1} dp =$$

$$= \frac{\alpha_j}{|\alpha|} \int_{\mathbb{S}_k} \frac{\Gamma(\alpha_1 + \dots + \beta_j + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\beta_j) \dots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \dots p_j^{\beta_j - 1} \dots p_j^{\alpha_k - 1} dp.$$

Ent nüüd märkame, et integreeritav funktsioon on $\mathrm{Dir}(k; \alpha_1, \dots, \beta_j, \dots, \alpha_k)$ jaotusega juhusliku vektori tihedusfunktsioon. Tihedusfunktsiooni omaduse kohaselt

$$E(\mathbf{p_j}) = \frac{\alpha_j}{|\alpha|} \int_{\mathbb{S}_k} \frac{\Gamma(\alpha_1 + \dots + \beta_j + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\beta_j) \dots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \dots p_j^{\beta_j - 1} \dots p_j^{\alpha_k - 1} dp = \frac{\alpha_j}{|\alpha|} \cdot 1 = \frac{\alpha_j}{|\alpha|}.$$

Niisiis,

$$E(\mathbf{p}) = (E(\mathbf{p_1}), \dots, E(\mathbf{p_k})) = \left(\frac{\alpha_1}{|\alpha|}, \dots, \frac{\alpha_k}{|\alpha|} \right).$$

□

Lause 2.10. ***Dispersioon.*** *Olgu* $\mathbf{p} = (\mathbf{p_1}, \dots, \mathbf{p_k})$ *juhuslik vektor jaotusega* $\mathit{Dir}(k; \alpha_1, \dots, \alpha_k)$. *Siis iga* $j \in \{1, \dots, k\}$ *korral juhusliku suuruse* $\mathbf{p_j}$ *dispersioon on*

$$D(\mathbf{p_j}) = \frac{\alpha_1}{|\alpha|^2} \cdot \frac{|\alpha| - \alpha_j}{|\alpha| + 1},$$

kus $|\alpha| = \sum_{i=1}^{k} \alpha_k$.

Tõestus. Olgu $j \in \{1, \dots, k\}$ suvaline. Dispersiooni definitsiooni kohaselt

$$D(\mathbf{p_j}) = E(\mathbf{p_j}^2) - (E(\mathbf{p_j}))^2.$$

Avaldame $E(\mathbf{p_j}^2)$. Keskväärtuse definitsioonist lähtuvalt

$$E(\mathbf{p_j}^2) = \int_{\mathbb{S}_k} p_j^2 \cdot \pi(p)dp = \int_{\mathbb{S}_k} p_j^2 \cdot \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \cdot \prod_{i=1}^{k} p_i^{\alpha_i - 1} dp =$$

$$= \int_{\mathbb{S}_k} \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \cdot \prod_{\substack{i=1,\\ i\neq j}}^{k} p_i^{\alpha_i - 1} \cdot p_j^{\alpha_i + 1} dp.$$

Tähistame $\beta_j = \alpha_j + 2$. Märkame, et gammafunktsiooni omaduse tõttu

$$\Gamma(\beta_j) = (\alpha_j + 1)\Gamma(\alpha_j + 1) = \alpha_j(\alpha_j + 1)\Gamma(\alpha_j),$$

$$\Gamma(|\alpha| + 2) = (|\alpha| + 1)\Gamma(|\alpha| + 1) = |\alpha|(|\alpha| + 1)\Gamma(|\alpha|).$$

Seega

$$\int_{\mathbb{S}_k} \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \cdot \prod_{\substack{i=1,\\ i\neq j}}^{k} p_i^{\alpha_i - 1} \cdot p_j^{\alpha_i + 1} dp =$$

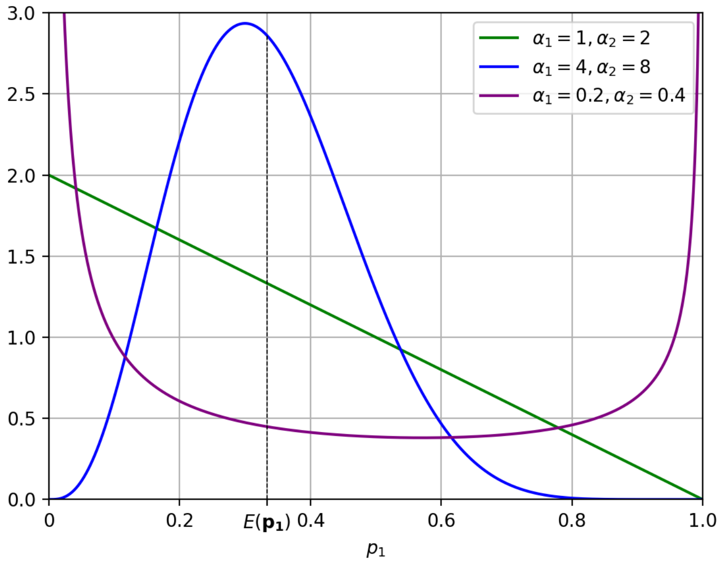
$$= \frac{\alpha_j(\alpha_j + 1)}{|\alpha|(|\alpha| + 1)} \int_{\mathbb{S}_k} \frac{\Gamma(|\alpha| + 2)}{\Gamma(\alpha_1)\ldots\Gamma(\beta_j)\ldots\Gamma(\alpha_k)} \cdot \prod_{\substack{i=1,\\ i\neq j}}^{k} p_i^{\alpha_i - 1} \cdot p_j^{\beta_j - 1} dp = \frac{\alpha_j(\alpha_j + 1)}{|\alpha|(|\alpha| + 1)}.$$

Märgime, et viimane võrdus kehtib, sest integraali alla jääb jaotuse $\mathrm{Dir}(k; \alpha_1, \ldots, \beta_j, \ldots, \alpha_k)$ tihedusfunktsioon, mistõttu on integraali väärtus 1.
Lause 2.9 kohaselt $E(\mathbf{p_j}) = \dfrac{\alpha_j}{|\alpha|}$. Niisiis,

$$D(\mathbf{p_j}) = \frac{\alpha_j(\alpha_j + 1)}{|\alpha|(|\alpha| + 1)} - \frac{\alpha_j^2}{|\alpha|^2} = \frac{\alpha_j}{|\alpha|}\left(\frac{\alpha_j + 1}{|\alpha| + 1} - \frac{\alpha_j}{|\alpha|}\right) = \frac{\alpha_j}{|\alpha|}\left(\frac{|\alpha| - \alpha_j}{|\alpha|(|\alpha| + 1)}\right) = \frac{\alpha_j}{|\alpha|^2}\left(\frac{|\alpha| - \alpha_j}{|\alpha| + 1}\right).$$

□

Vaatleme nüüd, millist rolli mängivad Dirichlet' jaotusega juhusliku vektori keskväärtus ja selle koordinaatide dispersioonid eeljaotuse parameetrite $\alpha_1, \ldots, \alpha_k$ valikul. Paneme tähele, et Dirichlet' jaotusega juhusliku vektori koordinaadi $\mathbf{p_i}$ dispersioon on seda väiksem, mida suurem on paramaeetrite $\alpha_1, \ldots, \alpha_k$ summa $|\alpha|$. Seda seost illustreerib joonis 2.1, millel on esitatud jaotustele $\mathrm{Be}(0{,}2; 0{,}4)$, $\mathrm{Be}(1; 2)$ ja $\mathrm{Be}(4; 8)$ vastavate tihedusfunktsioonide graafikud.
Märgime, et kõigi kolme jaotuse puhul $E(\mathbf{p}_1) = \dfrac{1}{3}$, see-eest juhusliku suuruse $\mathbf{p}_1$ dispersioon on erinev. Jaotuse $\mathrm{Be}(0{,}2; 0{,}4)$ korral $|\alpha| = 0{,}6$, mille tõttu koondub tõenäosusmass äärmuslike väärtuste 0 ja 1 juurde. Jaotuse $\mathrm{Be}(4; 8)$ puhul aga $|\alpha| = 12$ ning tõenäosusmass kontsntreerub keskväärtuse $E(\mathbf{p}_1)$ ümber. Jaotus $\mathrm{Be}(1; 2)$ on siinpuhul vahepealne variant, sest $|\alpha| = 3$ ja tõenäosusmass on seega ka ühtlasemalt jaotatud.



Joonis 2.1: Beetajaotuste tihedusfunktsioonid

Iga $i \in \{1, \ldots, k\}$ korral esitub parameeter α_i $\mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$ jaotusega vektori koordinaadi $\mathbf{p_i}$ keskväärtuse ja $|\alpha|$ korrutisena: $\alpha_i = \frac{\alpha_i}{|\alpha|} \cdot |\alpha| = E(\mathbf{p_i}) \cdot |\alpha|$. Niisiis, eeljaotuse parameetrite valiku korral kirjeldab $E(\mathbf{p_i})$ aprioorset arusaama juhusliku vektori $\mathbf{p}$ vastava koordinaadi keskväärtusest, suurus $|\alpha|$ aga väljendab hinnangu kindlusastet. Eelnevalt kirjeldatud Dirichlet'-kategoorilises protsessis vaatluste arvu n kasvades kasvab ka suurus $|\alpha|$ ehk $|\alpha| \to \infty$, kui $n \to \infty$. Sisuliselt tähendab see, et vaatluste lisandumisel koondub järeljäotuse tihedus vektori p tõelise väärtuse ümber.

2.2 Polya urn

Polya urni protsess aitab kujundada intuitiivset arusaama Dirichlet'-kategoorilisest protsessist. Järgnevalt esitame selle definitsiooni ning näitame, et parameetrite $\alpha_1, \ldots, \alpha_k$ naturaalarvuliste väärtuste korral langevad Dirichlet'-kategooriline ning Polya urni protsess kokku.

Definitsioon 2.11. Polya urn. Olgu urnis k erinevat värvi kuuli (antud juhul käsitleme värve arvväärtustena $1, \ldots, k$), kusjuures igat värvi $j \in \{1, \ldots, k\}$ on $\alpha_j \in \mathbb{N}$ tükki. Igal sammul tõmbame urnist juhuslikult ühe palli ja paneme koos sama värvi palliga tagasi. Tähistagu Y_i i-ndal sammul välja võetud kuuli värvi.

Teoreem 2.12. *Olgu $\alpha_1, \dots, \alpha_k$ mingid naturaalarvud ja $X_1, X_2, \dots$ Dirichlet'-kategoorili-ne protsess, s.t $X_1, X_2, \dots$ on sõltumatud juhuslikud suurused jaotusega p, kus p on juhusliku vektori $\mathbf{p}$ fikseeritud väärtus ja $\mathbf{p}$ on jaotusega $Dir(k; \alpha_1, \dots, \alpha_k)$. Olgu $Y_1, Y_2, \dots$ vastav Polya urni protsess. Siis iga $n \in \mathbb{N}$ puhul*

$$(X_1, \dots, X_n, X_{n+1}) \stackrel{d}{=} (Y_1, \dots, Y_n, Y_{n+1})$$

ehk need juhuslikud vektorid on sama jaotusega.

Tõestus. Olgu $i \in \{1, \dots, k\}$ suvaline. Esimesel tõmbamisel on urnis α_i $i-$värvi palli ja $|\alpha| = \sum_{i=1}^{k} \alpha_i$ palli kokku. Järelikult esimesel tõmbamisel i-värvi palli saamise tõenäosus on $P(Y_1 = i) = \frac{\alpha_i}{|\alpha|}$. Teisalt,

$$P(X_1 = i) = \int_{\mathbb{S}_k} P(X_1 = i | \mathbf{p} = p) \pi(p) dp = \int_{\mathbb{S}_k} p_i \cdot \pi(p) dp = E(\mathbf{p_i}) = \frac{\alpha_i}{|\alpha|}.$$

Niisiis, iga $i \in \{1, \dots, k\}$ korral

$$P(X_1 = i) = \frac{\alpha_i}{|\alpha|} = P(Y_1 = i).$$

Vaatleme teist sammu. Olgu $i_1, i_2 \in \{1, \dots, k\}$ suvalised. Milline on tõenäosus, et $Y_2 = i_2$, kui $Y_1 = i_1$? Iga $j \in \{1, \dots, k\}$ korral defineerime funktsiooni

$$I_j(i) = \begin{cases} 0, & \text{kui } i \neq j, \\ 1, & \text{kui } i = j. \end{cases}$$

Esimese tõmbamise järel pandi urnist tõmmatud pall tagasi koos veel ühe sama värvi palliga. Niisiis, teise tõmbamise eel on urnis $\alpha_{i_2} + I_{i_2}(i_1)$ i_2-värvi palli, kokku aga $|\alpha| + 1$ palli. Järelikult $P(Y_2 = i_2 | Y_1 = i_1) = \frac{\alpha_2 + I_{i_2}(i_1)}{|\alpha| + 1}$.

Nüüd uurime, millise tõenäosusega $X_2 = i_2$ tingimusel, et $X_1 = i_1$. Kõigepealt märgime, et

$$P(X_2 = i_2 | X_1 = i_1) = \int_{\mathbb{S}_k} P(X_2 = i_2, \mathbf{p} = p | X_1 = i_1) dp.$$

Bayesi teoreemi kohaselt

$$P(X_2 = i_2, \mathbf{p} = p | X_1 = i_1) = \pi(p | X_1 = i_1) \cdot P(X_2 = i_2 | \mathbf{p} = p, X_1 = i_1).$$

Et X_1 ja X_2 on sõltumatud, siis X_1 jaotus sõltub üksnes parameetrist p, seega

$$P(X_2 = i_2 | \mathbf{p} = p, X_1 = i_1) = P(X_2 = i_2 | \mathbf{p} = p).$$

Niisiis,

$$P(X_2 = i_2, \mathbf{p} = p | X_1 = i_1) = \int_{\mathbb{S}_k} P(X_2 = i_2 | \mathbf{p} = p) \pi(p | X_1 = i_1) dp =$$
$$= \int_{\mathbb{S}_k} p_{i_2} \pi(p | X_1 = i_1) = E(\mathbf{p_{i_2}} | X_1 = i_1).$$

Teoreemi 2.6 kohaselt on $\mathbf{p}|X_1 = i_1$ jaotusega $\mathrm{Dir}(k; \alpha_1, \ldots, \alpha_{i_1} + 1, \ldots, \alpha_k)$, seega lause 2.9 põhjal $E(\mathbf{p_{i_2}}) = \dfrac{\alpha_{i_2} + 1}{|\alpha| + 1}$, kui $i_2 = i_1$, ja $E(\mathbf{p_{i_2}}) = \dfrac{\alpha_{i_2}}{|\alpha| + 1}$, kui $i_1 \neq i_2$. Kokkuvõttes suvaliste $i_1, i_2 \in \{1, \ldots, k\}$ korral

$$P(X_2 = i_2 | X_1 = i_1) = \frac{\alpha_{i_2} + I_{i_2}(i_1)}{|\alpha| + 1} = P(Y_2 = i_2 | Y_1 = i_1).$$

Nüüd vaatleme sammu $n+1$. Olgu $i_1, \ldots i_n, i_{n+1} \in \{1, \ldots, k\}$ suvalised ja n esimesel sammul urnist tõmmatud kuulide värvid vastavalt $Y_1 = i_1, \ldots, Y_n = i_n$. Sellisel juhul on enne nr $n+1$ kuuli tõmbamist urni lisandunud $\sum_{j=1}^{n} I_{i_{n+1}}(i_j) =: N^n_{i_{n+1}}$ tükki i_{n+1}-värvi kuuli. Järelikult urnis on $\alpha_{i_{n+1}} + N^n_{i_{n+1}}$ i_{n+1}-värvi kuuli. Kokku aga vastavalt $|\alpha| + n$ kuuli. Järelikult i_{n+1}-värvi kuuli tõmbamise tõenäosus sammul $n + 1$ on

$$P(Y_{n+1} = i_{n+1} | Y_n = i_n, \ldots, Y_1 = i_1) = \frac{\alpha_{\breve{i}n+1} + N^n_{i_{n+1}}}{|\alpha| + n}.$$

Uurime, milline on tõenäosus, et $X_{n+1} = i_{n+1}$, kui $X_1 = i_1, \ldots, X_n = i_n$.

$$P(X_{n+1} = i_{n+1} | X_1 = i_1, \ldots, X_n = i_n) = \int_{\mathbb{S}_k} P(X_{n+1} = i_{n+1}, \mathbf{p} = p | X_1 = i_1, \ldots, X_n = i_n) dp =$$
$$= \int_{\mathbb{S}_k} \pi(p | X_1 = i_1, \ldots, X_n = i_n) P(X_{n+1} = i_{n+1} | \mathbf{p} = p, X_1 = i_1, \ldots, X_n = i_n) dp.$$

Et määratud p korral ei sõltu X_{n+1} jaotus juhuslikest suurustest $X_1, \ldots, X_n$, siis

$$\int_{\mathbb{S}_k} \pi(p | X_1 = i_1, \ldots, X_n = i_n) P(X_{n+1} = i_{n+1} | \mathbf{p} = p, X_1 = i_1, \ldots, X_n = i_n) dp =$$
$$= \int_{\mathbb{S}_k} \pi(p | X_1 = i_1, \ldots, X_n = i_n) P(X_{n+1} = i_{n+1} | \mathbf{p} = p) dp =$$
$$= \int_{\mathbb{S}_k} p_{i_{n+1}} \pi(p | X_1 = i_1, \ldots, X_n = i_n) dp = E(\mathbf{p_{i_{n+1}}} | X_1 = i_1, \ldots, X_n = i_n).$$

Teoreemi 2.6 kohaselt on $\mathbf{p}|X_1 = i_1, \ldots, X_n = i_n$ jaotusega $\mathrm{Dir}(k; \alpha_1 + N^n_1, \ldots, \alpha + N^n_k)$. Seega lause 2.9 põhjal $E(\mathbf{p_{i_{n+1}}}) = \dfrac{\alpha_{i_{n+1}} + N^n_{i_{n+1}}}{|\alpha| + n}$.

Kokkuvõttes näeme, et suvaliste $i_1, \ldots, i_n \in \{1, \ldots, k\}$ korral

$$P(X_{n+1} = i_{n+1} | X_1 = i_1, \ldots, X_n = i_n) = \frac{\alpha_{i_{n+1}} + N^n_{i_{n+1}}}{|\alpha| + n} = P(Y_{n+1} = i_{n+1} | Y_1 = i_1, \ldots, Y_n = i_n).$$

Järelikult $(X_1, \ldots, X_n, X_{n+1}) \overset{d}{=} (Y_1, \ldots, Y_n, Y_{n+1})$. $\square$

Märkus 2.13. Iga $j \in \{1, \dots, k\}$ korral tähistagu Q_j^n j–värvi pallide arvu urnis n tõmbamise järel ja olgu $\mathbf{p} \sim \mathrm{Dir}(k; \alpha_1, \dots, \alpha_k)$. Siis eri värvi pallide osakaalude vektorite $\left(\frac{Q_1^n}{n}, \dots, \frac{Q_k^n}{n}\right)$ jaotused P_{Q_n} koonduvad nõrgalt juhusliku vektori $\mathbf{p}$ jaotuseks $\mathrm{Dir}(k; \alpha)$. Selle kohta öeldakse ka, et juhuslikud vektorid $\left(\frac{Q_1^n}{n}, \dots, \frac{Q_k^n}{n}\right)$ koonduvad jaotuse järgi vektoriks $\mathbf{p}$ ja tähistatakse

$$\left(\frac{Q_1^n}{n}, \dots, \frac{Q_k^n}{n}\right) \xrightarrow{d} \mathrm{Dir}(k; \alpha_1, \dots, \alpha_k).$$

Kuna $\mathbf{p}$ ja $\left(\frac{Q_1^n}{n}, \dots, \frac{Q_k^n}{n}\right)$ väärtuste hulk $\mathbb{S}_k$ on separaabel, siis lausete 1.7 ja 1.8 põhjal on jaotuste nõrk koondumine samaväärne Prohhorovi meetrika järgi koondumisega ehk $\rho(P_{Q_n}, \mathrm{Dir}(k; \alpha)) \xrightarrow{n} 0$.

Tõepoolest, olgu $\mathbf{p}|\alpha, k \sim \mathrm{Dir}(k; \alpha)$ ja $X_1, \dots, X_n|\mathbf{p} = p \overset{\text{i.i.d}}{\sim} p$. Suvalise $i \in \{1, \dots, m\}$ ning $j \in \{1, \dots, k\}$ korral vaatleme juhuslikku suurust $I_j \circ X_i$. Märkame, et $\sum_{i=1}^{n} I_j \circ X_i$ loendab seerias $X_1, \dots, X_n$ väärtuse j saavutanud vaatluste arvu. Vaatleme juhuslikku vektorit $(R_1, \dots, R_k) = R \sim M(n; p)$. Paneme tähele, et $\sum_{i=1}^{n} I_j \circ X_i = R_j$. Järelikult

$$E\Big(\sum_{i=1}^{n} I_j \circ X_i\Big) = E(R_j) = np_j.$$

Tugeva suurte arvude seaduse kohaselt fikseeritud jaotuse p korral

$$\frac{\sum_{i=1}^{n} I_j \circ X_i}{n} \to p_j \quad \text{peaaegu kindlasti}$$

protsessis $n \to \infty$. Kirjeldatavas mudelis on aga p_j alati Dirichlet' jaotusega juhusliku vektori $\mathbf{p}$ koordinaadi $\mathbf{p_j}$ mingi realisatsioon. Sellest on võimalik järeldada, et üldisemalt

$$\left(\frac{\sum_{i=1}^{n} I_1 \circ X_i}{n}, \dots, \frac{\sum_{i=1}^{n} I_k \circ X_i}{n}\right) \to (\mathbf{p}_1, \dots, \mathbf{p_k}) \quad \text{peaaegu kindlasti,}$$

s.t Dirichlet' kategoorilise protsessi väärtuste osakaalude vektor koondub peaaegu kindlasti Dirichlet' jaotusega juhuslikuks vektoriks. Märgime, et jaotuse järgi koondumine järeldub peaaegu kindlast koondumisest ja Polya urni ning Dirichlet' kategooriline protsess on iga $n \in \mathbb{N}$ korral sama jaotusega. Sellest saame järeldada, et Polya urni protsessi korral koondub värvide osakaalude vektorite jaotus Dirichlet' jaotuseks ehk

$$\left(\frac{Q_1^n}{n}, \dots, \frac{Q_k^n}{n}\right) \xrightarrow{d} \mathrm{Dir}(k; \alpha_1, \dots, \alpha_k).$$

2.3 Tokimurdmine

$\mathrm{Dir}(k;\alpha_1,\ldots,\alpha_k)$ jaotusega juhusliku vektori $\mathbf{p} = (\mathbf{p}_1,\ldots,\mathbf{p}_\mathbf{k})$ genereerimiseks saab kasutada nn tokimurdmise meetodit. Meetodi nimi tuleneb sellest, et intuitiivselt sarnaneb algoritm ühikpikkusega toki murdmisele k osaks. Järgnevalt esitame kokkuvõtte antud juhusliku vektori genereerimise viisist.

- Olgu $V_1 \sim \mathrm{Be}(\alpha_1, \sum_{i=2}^{k}\alpha_i)$. Võtame selle juhusliku vektori $\mathbf{p}$ esimeseks koordinaadiks $\mathbf{p}_1 = V_1$. Sisuliselt sarnaneb see tokist V_1-osa ära murdmisega (nt kui $V_1 = \frac{1}{4}$, murrame tokist ühe neljandiku).

- Teisel sammul võtame juhusliku suuruse $V_2 \sim \mathrm{Be}(\alpha_2, \sum_{i=3}^{k}\alpha_i)$ ja määrame vektori $\mathbf{p}$ teiseks koordinaadiks $\mathbf{p}_2 = V_2(1-V_1)$. Põhimõtteliselt kujutab see endast allesjäänud toki küljest V_2-osa murdmist.

- Iga $j \in \{3,\ldots,k-1\}$ korral võtame juhusliku suuruse $V_j \sim \mathrm{Be}(\alpha_2, \sum_{i=j+1}^{k})$ ja määrame vektori $\mathbf{p}$ j–ndaks koordinaadiks $\mathbf{p}_\mathbf{j} = V_j\prod_{i=1}^{j-1}(1 - V_i)$. Nagu ka eelnevate sammude puhul on see sisuliselt $j-1$ murdmise järel alles jäänud tokist V_j-osa murdmine.

- Vektori $\mathbf{p}$ k–ndaks koordinaadiks määrame $k-1$ murdmise järel alles jäänud osa.

On võimalik näidata, et juhul, kui juhuslikud suurused $V_1,\ldots,V_k$ on sõltumatud, siis niisugusel viisil genereeritud juhuslik vektor on tõepoolest jaotusega $\mathrm{Dir}(k;\alpha_1,\ldots,\alpha_k)$. Viimase fakti tõestuse leiab näiteks [5, lk 564–565].

3. Dirichlet' protsess

Märkus 3.1. Olgu $(\Omega, \mathcal{F}, \mathbf{P})$ mingi tõenäosusruum. Mõiste juhuslik protsess tähistab sellisel juhul ruumil $(\Omega, \mathcal{F}, \mathbf{P})$ määratud juhuslike suuruste hulka $\{X_t\}_{t \in T}$, kus T on omakorda suvaline indeksite hulk. Mis õigustab nimetust "Dirichlet' protsess"– tegemist on ju teatud omadust rahuldava juhusliku tõenäosusmõõduga (juhusliku suurusega) $P : \Omega \to \mathcal{P}$? Märgime, et võttes $T = \mathcal{B}(S)$ ja iga $A \in \mathcal{B}(S)$ korral $t = A$, saame defineerida juhusliku protsessi $\{X_t\}_{t \in T}$, kus $X_t : \Omega \to [0,1]$, $\omega \mapsto P(\omega)(A)$. Niisiis, saame vaadelda igat juhuslikku mõõtu juhusliku protsessina.

Kujutame ette olukorda, kus on antud teatud vaatlused $X_1 = x_1, \ldots, X_n = x_n$, kuid puudub info selle kohta, mis jaotusest p need pärit on. Sellisel juhul saab käsitleda vaatluste jaotust juhusliku suurusena (juhusliku tõenäosusmõõduna) P ning võtta Dirichlet' protsessi jaotuse mõõdu P eeljaotuseks. Intuitiivselt tähendab viimane seda, et kasutame Dirichlet' protsessi jaotust kirjeldamaks oma esialgset arusaama P jaotusest. Sisuliselt haakub niisugune probleemipüstitus eelmises peatükis kirjeldatud Dirichlet'-kategoorilise protsessi mudeliga, kuid enam ei eeldata, et $p \in \mathbb{S}_k$, vaid p saab olla suvaline diskreetne tõenäosusmõõt (hiljem selgitame, miks diskreetsuse nõue ei ole praktikas kuigi kitsendav).

Üks Dirichlet' protsessi jaotuse suurimaid voorusi on see, et äsja kirjeldatud olukorras osutub ka järeljaotus Dirichlet' protsessi jaotuseks. Lihtsamalt öeldes, kui uuendame olemasolevate andmete valguses oma arusaama P jaotusest, siis selle üldine kuju ei muutu. Veelgi enam: vaatluste arvu n kasvades koondub P järeljaotus väärtuseks p ehk nende tõeliseks jaotuseks. Ka siis, kui esialgne arusaam jaotusest p on vale, saame piisava hulga (empiiriliste) andmete korral selle tõelisele väärtusele lähedale.

Eelkirjeldatu teeb Dirichlet' protsessi jaotusest asendamatu töövahendi mitteparameetrilises Bayesi statistikas. Selles peatükis esitame Dirichlet' protsessi definitsiooni ja loome raamistiku selle olemasolu tõestamiseks. Seejärel uurime Dirichlet' protsessist tehtud vaatluste mudelit ja lõpuks esitame skeemid Dirichlet' prosessi ja vastava juhusliku valimi genereerimiseks. Peatükis esitatavad tulemused tuginevad [5, lk 59–69], kui ei ole märgitud teisiti.

Märkus 3.2. Tuletame meelde mõningad objektid esimesest peatükist.

- Olgu (S,d) Poola ruum ja $\mathcal{B}(S)$ vastav Boreli σ-algebra.
- Tähistagu $\mathcal{P}$ kõigi ruumil $(S,\mathcal{B}(S))$ defineeritud tõenäosusmõõtude hulka. Esimeses peatükis veendusime, et $(\mathcal{P},\rho)$, kus ρ tähistab Prohhorovi meetrikat, on samuti Poola ruum (vt teoreeme 1.9 ja 1.15).
- Et ρ on kaugus, siis indutseerib see topoloogia τ_ρ. Niisiis saame rääkida ka ruumi $\mathcal{P}$ Boreli σ-algebrast $\mathcal{B}(P) = \sigma(\tau_\rho)$.
- Suvalise $A \in \mathcal{B}(S)$ korral tähistab T_A kujutust $T_A : \mathcal{P} \to [0,1]$, mis on määratud seosega $T_A(p) = p(A)$.

Selles peatüki ulatuses kasutame eelmainitud tähiseid äsja kirjeldatud tähenduses.

3.1 Definitsioon ja olemasolu

Definitsioon 3.3. Juhuslik tõenäosusmõõt. Olgu $(\Omega,\mathcal{F},\mathbf{P})$ mingi tõenäosusruum. Juhuslikuks tõenäosusmõõduks ruumil $(S,\mathcal{B}(S))$ nimetatakse $\mathcal{F}$-mõõtuvat funktsiooni $P : \Omega \to \mathcal{P}$. Kujutuse P mõõtuvus tähendab antud juhul seda, et iga mõõtuva hulga $E \subset \mathcal{P}$ korral $P^{-1}(E) \in \mathcal{F}$.

Märkus 3.4. Märgime, et iga $A \in \mathcal{B}(S)$ korral saame vaadelda kompositsiooni

$$T_A \circ P : \Omega \to [0,1], \quad \omega \mapsto (P(\omega))(A).$$

Kuna T_A on teoreemi 1.16 põhjal $\mathcal{B}(\mathcal{P})$-mõõtuv ja P on $\mathcal{F}$-mõõtuv, siis ka nende kompositsioon on mõõtuv ehk samuti juhuslik suurus. Edasipidi tähistame kompositsiooni $T_A \circ P$ lihtsalt kui $P(A)$.

Definitsioon 3.5. Dirichlet' protsess. Olgu α lõplik mõõt ruumil $(S, \mathcal{B}(S))$. Dirichlet' protsessiks baasimõõduga α nimetatakse juhuslikku tõenäosusmõõtu P ruumil $(S, \mathcal{B}(S))$, mis rahuldab ruumi S iga lõpliku mõõtuva tükelduse $A_1, \ldots, A_k$ korral omadust

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)).$$

Definitsioon 3.6. Dirichlet' protsessi jaotus. Olgu $(\Omega, \mathcal{F}, \mathbf{P})$ mingi tõenäosusruum ja $P : \Omega \to \mathcal{P}$ Dirichlet' protsess baasimõõduga α. P määrab tõnäosusmõõdu Π^* ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ seosega:

$$\Pi^*(E) := \mathbf{P}P^{-1}(E), \quad \forall E \in \mathcal{B}(\mathcal{P}).$$

Seejuures mõõtu Π^* nimetatakse Dirichlet' protsessi P jaotuseks.

Definitsioon 3.7. Tõenäosusmõõt $\mathbf{\Pi}(\alpha)$. Olgu α lõplik tõenäosusmõõt ruumil $(S, \mathcal{B}(S))$. Mõõduks $\Pi(\alpha)$ nimetame sellist tõenäosusmõõtu ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, mille korral

$$(T_{A_1}, \ldots, T_{A_k}) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k))$$

ruumi S iga mõõtuva tükelduse $A_1, \ldots, A_k$ puhul.

Järgnev teoreem võimaldab taandada Dirichlet' protsessi olemasolu mõõdu $\Pi(\alpha)$ olemasolule.

Teoreem 3.8. *Tõenäosusmõõt Π ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ on Dirichlet' protsessi P jaotus parajasti siis, kui $\Pi = \Pi(\alpha)$.*

Tõestus. **Tarvikkus.** Eeldame, et Π on Dirichlet' protsessi P jaotus. Olgu $A_1, \ldots, A_k$ ruumi S suvaline mõõtuv tükeldus. Lause 1.16 põhjal iga $j \in \{1, \ldots, k\}$ korral kujutus $T_{A_j} : \mathcal{P} \to [0, 1]$ on $\mathcal{B}(\mathcal{P})$-mõõtuv, seega iga T_{A_j} on juhuslik suurus ja

$$(T_{A_1}, \ldots, T_{A_k}) : \mathcal{P} \to \mathbb{S}_k$$

juhuslik vektor.

Kuna Π on tõenäosusmõõt ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, siis tõenäosusmõõt $\Pi(T_{A_1}, \ldots, T_{A_k})^{-1}$ ruumil $(\mathbb{S}_k, \mathcal{B}(\mathbb{S}_k))$ on juhusliku vektori $(T_{A_1}, \ldots, T_{A_k})$ jaotus. Et aga Π on ka Dirichlet' protsessi jaotus, siis

$$\Pi(T_{A_1}, \ldots, T_{A_k})^{-1} = \mathbf{P}P^{-1}(T_{A_1}, \ldots, T_{A_k})^{-1}.$$

Nüüd aga paneme tähele, et suvalise $B \in \mathcal{B}(\mathbb{S}_k)$ korral

$$P^{-1}(T_{A_1}, \ldots, T_{A_k})^{-1}(B) = ((T_{A_1}, \ldots, T_{A_k}) \circ P)^{-1}(B) =$$
$$= (T_{A_1} \circ P, \ldots, T_{A_k} \circ P)^{-1}(B) = (P(A_1), \ldots, P(A_k))^{-1}(B).$$

Järelikult $\Pi(T_{A_1}, \ldots, T_{A_k})^{-1}$ on juhusliku vektori $(P(A_1), \ldots, P(A_k))$ jaotus. Niisiis,

$$(T_{A_1}, \ldots, T_{A_k}) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)).$$

Näeme, et mõõt Π rahuldab definitsiooni 3.7, seega $\Pi = \Pi(\alpha)$.

Piisavus. Segaduse vältimiseks tähistame tõenäosusmõõtu $\Pi(\alpha)$ selle tõestuse raames sümboliga Π_α. Eeldame, et $\Pi = \Pi_\alpha$. Veendume, et Π on Dirichlet' protsessi jaotus. Vaatleme tõenäosusruumi $(\mathcal{P}, \mathcal{B}(\mathcal{P}), \Pi)$ ja juhuslikku tõenäosusmõõtu $P : \mathcal{P} \to \mathcal{P}$, $p \mapsto p$. Siis iga $E \in \mathcal{B}(\mathcal{P})$ korral

$$\Pi(E) = \Pi P^{-1}(E) = \Pi_\alpha P^{-1}(E) = \Pi_\alpha(E)$$

ehk $\Pi = \Pi_\alpha$ on juhusliku tõenäosusmõõdu P jaotus. Teisalt iga $p \in \mathcal{P}$ korral

$$(P(A_1), \ldots, P(A_k))(p) = (T_{A_1} \circ P, \ldots, T_{A_k} \circ P)(p) =$$
$$= (T_{A_1}(P(p)), \ldots, T_{A_k}(P(p))) = (T_{A_1}(p), \ldots, T_{A_k}(p)) = (T_{A_1}, \ldots, T_{A_k})(p).$$

Niisiis, eelduse kohaselt mõõdu Π puhul

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)$$

ehk P on Dirichlet' protsess. Järelikult Π on Dirichlet' protsessi jaotus. □

Lühidalt kokku võttes väidab teoreem 3.8, et Dirichlet' protsess eksisteerib parajasti siis, kui leidub tõenäosusmõõt $\Pi(\alpha)$ ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, et hulga S iga lõpliku mõõtuva tükelduse $A_1, \ldots, A_k$ korral

$$(T_{A_1}, \ldots, T_{A_k}) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)).$$

Lause 3.9. *Kui eksisteerib definitsiooni 3.7 rahuldav tõenäosusmõõt, siis see on ühene.*

Tõestus. Olgu α lõplik mõõt ruumil $(S, \mathcal{B}(S))$ ning Π ja Π' kaks tõenäosusmõõtu ruumil $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, kusjuures kumbki rahuldab definitsiooni 3.7. Olgu $A_1, \ldots, A_k$ ruumi S suvaline mõõtuv tükeldus. Eelduse kohaselt juhusliku vektori $(T_{A_1}, \ldots, T_{A_k})$ jaotused

$\Pi(T_{A_1}, \dots, T_{A_k})^{-1}$ ja $\Pi'(T_{A_1}, \dots, T_{A_k})^{-1}$ ühtivad (mõlemad on Dirichlet' jaotused parameetritega $\alpha(A_1), \dots, \alpha(A_k)$). See tähendab, et suvalise $B \in \mathcal{B}(\mathbb{S}_k)$ korral

$$\Pi\{p \in \mathcal{P} : (T_{A_1}(p), \dots, T_{A_k}(p)) \in B\} = \Pi'\{p \in \mathcal{P} : (T_{A_1}(p), \dots, T_{A_k}(p)) \in B\}.$$

Teoreemi 1.19 põhjal järeldub siit, et $\Pi = \Pi' = \Pi(\alpha)$. □

Kui mõõt $\Pi(\alpha)$ eksisteerib, siis teoreemi 3.9 põhjal on see ka ühene. Mõõdu $\Pi(\alpha)$ olemasolu saab tõestada Kolmogorovi olemasolu teoreemi kaudu. Niisugune tõestus on esitatud näiteks [5, lk 69–70]. Asjaolu, et juhuslik mõõt P on Dirichlet' protsess baasimõõduga α, tähistame edaspidi $P \sim \Pi(\alpha)$.

Lõpetuseks vaatleme veel ühte loomulikku küsimust, mis Dirichlet' protsessi defineerimisel tekib. Kui $P \sim \Pi(\alpha)$, siis definitsioonist 3.5 lähtuvalt

$$(P(A_1), \dots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \dots, \alpha(A_k))$$

ruumi S iga lõpliku mõõtuva tükelduse $A_1, \dots, A_k$ korral. Olgu $A_1, \dots, A_k$ niisugune ruumi S tükeldus. Siis ka $A_1 \cup A_2, A_3, \dots, A_k$ on lõplik mõõtuv tükeldus. Kui $(P(A_1), \dots, P(A_k))$ on jaotusega $\mathrm{Dir}(k; \alpha(A_1), \dots, \alpha(A_k))$, siis mis garanteerib selle, et $(k-1)$-mõõtmeline vektor $(P(A_1 \cup A_2), P(A_3), \dots, P(A_k))$ on jaotusega $\mathrm{Dir}(k-1; \alpha(A_1 \cup A_2), \alpha(A_3), \dots, \alpha(A_k))$? Antud juhul tulevad appi baasimõõdu α aditiivsus ja teoreem 2.4.
Kõigepealt paneme tähele, et suvalise $p \in \mathcal{P}$ korral

$$\begin{aligned} T_{A_1 \cup A_2}(p) &= p(A_1 \cup A_2) = p(A_1) + p(A_2) = \\ &= T_{A_1}(p) + T_{A_2}(p) = (T_{A_1} + T_{A_2})(p). \end{aligned}$$

Järelikult

$$P(A_1 \cup A_2) = (T_{A_1} + T_{A_2}) \circ P = T_{A_1} \circ P + T_{A_2} \circ P = P(A_1) + P(A_2).$$

Niisiis,

$$(P(A_1 \cup A_2), P(A_3), \dots, P(A_k)) = (P(A_1) + P(A_2), P(A_3), \dots, P(A_k)).$$

Teoreemi 2.4 põhjal

$$(P(A_1) + P(A_2), P(A_3), \dots, P(A_k)) \sim \mathrm{Dir}(k-1; \alpha(A_1) + \alpha(A_2), \alpha(A_3), \dots, \alpha(A_k)).$$

Ent kuna A_1 ja A_2 on lõikumatud, siis $\alpha(A_1) + \alpha(A_2) = \alpha(A_1 \cup A_2)$. Järelikult

$$(P(A_1 \cup A_2), P(A_3), \dots, P(A_k)) \sim \mathrm{Dir}(k-1; \alpha(A_1 \cup A_2), \alpha(A_3), \dots, \alpha(A_k)).$$

Kui α on lõplik mõõt ruumil $(S, \mathcal{B}(S))$, siis kasutame tähistusi $M = \alpha(S)$, $\bar{\alpha} = \dfrac{\alpha}{M}$. Seejuures märgime, et kirjapilt $\alpha = M\bar{\alpha}$ on intuitiivselt mõneti loogilisem, sest praktikas valitakse enamasti eraldi tõenäosusmõõt $\bar{\alpha}$ ja kontsentratsiooniparameeter M, mõõt α määratakse aga nende kaudu. Selgitamaks, mis loogika järgi seda valikut tehakse, vaatleme kõigepealt juhusliku suuruse $P(A)$ keskväärtust ja dispersiooni. Seejärel defineerime veel vaatlused Dirichlet' protsessist.

Definitsioon 3.10. Keskväärtus. Olgu $P \sim \Pi(\alpha)$. Iga $A \in \mathcal{B}(S)$ on juhusliku suuruse $P(A)$ keskväärtus defineeritud järgneva seosega:

$$E(P(A)) := \int_\Omega P(\omega)(A) d\mathbf{P}(\omega) = \int_\Omega T_A(P(\omega)) d\mathbf{P}(\omega) = \int_\mathcal{P} T_A(p) d\Pi(p) = \int_\mathcal{P} p(A) d\Pi(p).$$

Lause 3.11. *Olgu P juhuslik tõenäosusmõõt ruumil $(S, \mathcal{B}(S))$ jaotusega $\Pi(\alpha)$. Siis suvalise $A \in \mathcal{B}(S)$ korral*

$$E(P(A)) = \bar{\alpha}(A).$$

Tõestus. Olgu $A \subset S$ suvaline mõõtuv hulk. Siis ka A^c on mõõtuv ja $\{A, A^c\}$ on ruumi S mõõtuv tükeldus. Kuna P on $\Pi(\alpha)$ jaotusega juhuslik tõenäosusmõõt, siis

$$(P(A),\ P(A^c)) \sim \mathrm{Dir}(2; \alpha(A), \alpha(A^c)).$$

Lause 2.9 kohaselt

$$E(P(A)) = \frac{\alpha(A)}{\alpha(A) + \alpha(A^c)}.$$

Kuna A ja A^c on lõikumatud, siis

$$\frac{\alpha(A)}{\alpha(A) + \alpha(A^c)} = \frac{\alpha(A)}{\alpha(A \cup A^c)} = \frac{\alpha(A)}{\alpha(S)} = \frac{\alpha(A)}{M} = \bar{\alpha}(A).$$

$\square$

Lause 3.12. *Dispersioon.* *Olgu P juhuslik tõenäosusmõõt ruumil $(S, \mathcal{B}(S))$ jaotusega $\Pi(\alpha)$. Siis suvalise $A \in \mathcal{B}(S)$ korral*

$$D(P(A)) = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1 + M}.$$

Tõestus. Olgu $A \subset S$ suvaline mõõtuv hulk. Siis ka A^c on mõõtuv ja $\{A, A^c\}$ on ruumi S mõõtuv tükeldus. Kuna P on $\Pi(\alpha)$ jaotusega juhuslik tõenäosusmõõt, siis

$$(P(A),\ P(A^c)) \sim \mathrm{Dir}(2; \alpha(A), \alpha(A^c)).$$

Lause 2.10 kohaselt

$$D(P(A)) = \frac{\alpha(A)}{((\alpha(A) + \alpha(A^c))^2} \cdot \frac{(\alpha(A) + \alpha(A^c)) - \alpha(A)}{(\alpha(A) + \alpha(A^c)) + 1} =$$

$$= \frac{\alpha(A)}{\alpha^2(A \cup A^c)} \cdot \frac{\alpha(A \cup A^c) - \alpha(A)}{\alpha(A \cup A^c) + 1} = \frac{\alpha(A)}{\alpha^2(S)} \cdot \frac{\alpha(S) - \alpha(A)}{\alpha(S) + 1} =$$

$$= \frac{\alpha(A)}{M^2} \cdot \frac{M - \alpha(A)}{M + 1} = \frac{\bar{\alpha}(A)}{M} \cdot \frac{M(1 - \frac{\alpha(A)}{M})}{1 + M} = \frac{\bar{\alpha}(A)(1 - \bar{\alpha}(A))}{1 + M} = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1 + M}.$$

□

Definitsioon 3.13. Vaatlused Dirichlet' protsessist. Olgu α lõplik mõõt ruumil $(S, \mathcal{B}(S))$. Vaatleme juhuslikku vektorit $(P, X_1, \ldots, X_n)$, kus $P \sim \Pi(\alpha)$ ja juhusliku tõenäosusmõõdu P fikseeritud väärtuse p korral $X_1, \ldots, X_n$ sõltumatud juhuslikud suurused jaotusega p, mida nimetame vaatlusteks Dirichlet' protsessist. Äsja kirjelatud mudelit tähistame järgmiselt

$$P|\alpha \sim \Pi(\alpha),$$
$$X_1, \ldots, X_n|P = p \overset{\text{i.i.d}}{\sim} p.$$

Äsja defineeritud mudeli ja lausete 3.11 ning 3.12 valguses tuleme tagasi tõenäosusmõõdu $\bar{\alpha}$ ja kontsentratsiooniparameetri M valiku loogika juurde. Olgu meil vaatlused $X_1 = x_1, \ldots X_n = x_n$, kusjuures me arvame, et tegemist on valimiga jaotusest $N(0;1)$ (normaaljaotus keskväärtusega 0 ja dispersiooniga 1), aga ei ole oma hinnagus kindlad. Sellisel juhul saame vaadata juhuslikke suurusi $X_1, \ldots, X_n$ vaatlustena Dirichlet' protsessit baasimõõduga $M \cdot N(0;1)$.

Märgime etteruttavalt, et iga $j \in \{1, \ldots, n\}$ korral $P(X_j \in A) = E(P(A)) = \bar{\alpha}(A)$ (seda väidet põhjendame teoreemis 3.17). Niisiis, tõenäosusmõõt $\bar{\alpha}$ postuleerib esialgse arusaama suuruste $X_1, \ldots, X_n$ jaotusest. Kirjeldatud probleemi juures püstitasime hüpoteesi, et juhuslike suuruste jaotus on $N(0;1)$, seega valime $\bar{\alpha} = N(0;1)$.

Pöördume nüüd parameetri M valiku juurde. Lause 3.12 kohaselt on $P(A)$ dispersioon pöördvõrdelises seoses kontsentratsiooniparameetriga M. Mida suurem on arv M, seda tugevamini kontsentreerub tõenäosusmass juhusliku mõõdu P eeljaotuses $\bar{\alpha}$ ümber. Seega intuitiivselt näitab M, kui tugevalt me postuleeritud hüpoteesi usume.

Kui $P|\alpha \sim \Pi(\alpha)$ ja $X_1, \ldots, X_n|P = p \overset{\text{i.i.d}}{\sim} p$, siis $\Pi(\alpha)$ on juhuslike suuruste $X_1, \ldots, X_n$ jaotuse eeljaotus, s.t kujutab endast arusaama juhuslikust mõõdust P enne andmete $x_1, \ldots, x_n$ arvesse võtmist. Üldjuhul on selline esialgne arusaam ebatäpne. Nagu ka Dirichlet'-kategoorilise protsessi puhul, pakub Dirichlet' protsessist tehtud vaatluste korral huvi, milline on juhusliku tõenäosusmõõdu P järeljaotus ehk milline on tinglik jaotus $P|X_1 = x_1, \ldots X_n = x_n$ antud andmete $x_1, \ldots, x_n \in S$ korral. Teoreemis 3.15 tõestame, et ka järeljaotus on Dirichlet' protsessi jaotus. Eelnevalt aga esitame tõestuseks vajaliku abitulemuse.

[5, lk 61–62] on näidatud, et Dirichlet' protsess on sabavaba (*tail-free*) protsess. Me ei peatu pikalt selle mõiste tähendusel, märgime vaid, et definitsioon on esitatud [5, lk 40] ja seotud mõisteid käsitletatakse põhjalikult [5, lk 37–39]. Kuna Dirichlet' protsess on sabavaba, siis [5, lk 42] tõestatud tulemuse põhjal formuleerime järgneva lause.

Lause 3.14. *Olgu $A_1, \ldots, A_k$ ruumi S mõõtuv tükeldus,*

$$P|\alpha \sim \Pi(\alpha),$$
$$X_1, \ldots, X_n|P = p \overset{i.i.d}{\sim} p$$

Tähistagu iga $j \in \{1, \ldots, k\}$ korral $R_j = \sum_{i=1}^{n} \delta_{X_i}$. Siis juhusliku vektori $(P(A_1), \ldots, P(A_k))$ järeljaotused

$$(P(A_1), \ldots, P(A_k))|X_1 = x_1, \ldots, X_n = x_n$$

ja

$$(P(A_1), \ldots, P(A_k))|R_1 = r_1, \ldots, R_k = r_k$$

langevad kokku peaaegu iga $X_1, \ldots, X_n$ realisatsiooni $x_1, \ldots, x_n$ korral.

Teoreem 3.15. *Olgu $A_1, \ldots, A_k$ mingi ruumi S lõplik mõõtuv tükeldus, $P|\alpha \sim \Pi(\alpha)$ ja $X_1, \ldots, X_n|P = p \overset{i.i.d}{\sim} p$. Siis peaaegu iga $X_1, \ldots, X_n$ realisatsiooni $x_1, \ldots, x_n$ korral*

$$P|X_1 = x_1, \ldots, X_n = x_n \sim \Pi\Big(\alpha + \sum_{i=1}^{n} \delta_{x_i}\Big).$$

Tõestus. Olgu $A_1, \dots, A_k$ ruumi S suvaline lõplik mõõtuv tükeldus ja $x_1, \dots, x_n$ selline $X_1, \dots, X_n$ realisatsioon, mille korral kehtib lause 3.14. Kuna $P \sim \Pi(\alpha)$, siis

$$(P(A_1), \dots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \dots, \alpha(A_k)).$$

Iga $j \in \{1, \dots, k\}$ korral olgu $r_j = \sum_{i=1}^{n} \delta_{x_i}(A_j)$ ja $R_j = \sum_{i=1}^{n} \delta_{X_i}(A_j)$. Tõenäosus, et X_i kuulub hulka A_j on $p(A_j)$, kus $(p(A_1), \dots, p(A_k))$ on vektori $(P(A_1), \dots, P(A_k))$ fikseeritud väärtus. Iga R_j on juhuslik suurus, mis loendab n vaatluse seas hulka A_j sattunute arvu ehk siis $(R_1, \dots, R_k) \sim M(n; p(A_1), \dots, p(A_k))$. Lause 3.14 põhjal ühtib järeljaotus

$$(P(A_1), \dots, P(A_k)) | X_1 = x_1, \dots, X_n = x_n$$

järeljaotusega

$$(P(A_1), \dots, P(A_k)) | R_1 = r_1, \dots, R_k = r_k.$$

Järelduse 2.8 kohaselt aga

$$(P(A_1), \dots, P(A_k)) | R_1 = r_1, \dots, R_k = r_k \sim \mathrm{Dir}(k; \alpha(A_1) + r_1, \dots, \alpha(A_k) + r_k).$$

Paneme tähele, et iga $j \in \{1, \dots, k\}$ korral

$$\alpha(A_j) + r_j = \alpha(A_j) + \sum_{i=1}^{n} \delta_{x_i}(A_j) = \Big(\alpha + \sum_{i=1}^{n} \delta_{x_i}\Big)(A_j).$$

Et eelnev kehtib iga tükelduse $A_1, \dots, A_k$ korral, siis P järeljaotus on $\Pi\Big(\alpha + \sum_{i=1}^{n} \delta_{x_i}\Big)$ peaaegu iga $X_1, \dots, X_n$ realisatsiooni $x_1, \dots, x_n$ puhul. $\square$

3.2 Üldistatud Polya urni protsess

Olgu $X_1, \dots, X_{n+1}$ vaatlused Dirichlet' protsessist. Nende ühisjaotusel on tihtipeale keerukas kuju, kuid seda saab kergesti kirjeldada tinglike jaotuste $X_1, X_2 | X_1, \dots, X_{n+1} | X_1, \dots, X_n$ kaudu [5, lk 64]. Anname sellele mõttele põgusa illustratsiooni. Olgu $A_1, \dots, A_k$ ruumi S mõõtuv tükeldus ja $i_1, \dots, i_{n+1} \in \{1, \dots, k\}$. Siis tingliku tõenäosuse valemi järjestikusel rakendamisel saame

$$P(X_1 \in A_{i_1}, \dots, X_{n+1} \in A_{i_{n+1}}) =$$
$$= P(X_{n+1} \in A_{i_{n+1}} | X_n \in A_{i_n}, \dots, X_1 \in A_{i_1}) \dots P(X_2 \in A_{i_2} | X_1 \in A_{i_1}) P(X_1 \in A_{i_1}).$$

Seejuures osutub, et vastavaid tinglikke jaotusi saab hõlpsasti kirjeldada üldistatud Polya urni protsessi kaudu. Järgnevalt esitame selle definitsiooni ja tõestame, et üldistatud Polya urni protsess ja vaatlused Dirichlet' protsessist langevad kokku.

Definitsioon 3.16. Olgu $\bar{\alpha}$ lõplik tõenäosusmõõt ruumil $(S, \mathcal{B}(S))$. Juhuslike suuruste jada $Y_1, Y_2, \ldots$ nimetatakse üldistatud Polya urniks, kui $Y_1 \sim \bar{\alpha}$ ja iga $n \in \mathbb{N}$ korral

$$Y_{n+1}|Y_1, \ldots, Y_n \sim \begin{cases} \delta_{Y_1} & \text{tõenäosusega } 1/(M+n), \\ \vdots & \\ \delta_{Y_n} & \text{tõenäosusega } 1/(M+n), \\ \bar{\alpha} & \text{tõenäosusega } M/(M+n). \end{cases}$$

Teoreem 3.17. *Olgu $P \sim \Pi(\alpha)$ ja $X_1, X_2, \ldots \overset{i.i.d}{\sim} p$. Olgu $Y_1, Y_2, \ldots$ üldistatud Polya urni protsess. Siis iga $n \in \mathbb{N}$ korral vektorid $(X_1, \ldots, X_n, X_{n+1})$ ja $(Y_1, \ldots, Y_n, Y_{n+1})$ on sama jaotusega.*

Tõestus. Olgu A ruumi S suvaline mõõtuv alamhulk. Üldistatud Polya urni definitsiooni kohaselt on Y_1 juhuslik suurus jaotusega $\bar{\alpha}$. Järelikult tõenäosus, et $Y_1 \in A$, on $\bar{\alpha}(A)$. Vaatleme, milline on tõenäosus, et $X_1 \in A_{i_1}$.

$$P(X_1 \in A) = \int_\Omega P(\omega)(A) d\mathbf{P}(\omega) = \int_{\mathcal{P}} p(A) d\Pi(p) = E(P(A)) = \bar{\alpha}(A) \tag{3.1}$$

Nüüd fikseerime vabalt $n \in \mathbb{N}$. Olgu $x_1, \ldots, x_n \in S$ juhuslike suuruste $X_1, \ldots, X_n$ realisatsioon. Tõenäosus, et $Y_{n+1} \in A$ tingimusel, et $Y_1 = x_1, \ldots, Y_n = x_n$, on üldistatud Polya urni defintsiooni kohasel

$$\frac{M}{M+n}\bar{\alpha}(A) + \frac{1}{M+n}\sum_{i=1}^{n} \delta_{x_i}(A).$$

Leiame tõenäosuse $P(X_{n+1} \in A | X_1 = x_1, \ldots, X_n = x_n)$. Kuna $X_1, \ldots, X_n, X_{n+1}$ on vaatlused Dirichlet' protsessist, siis teoreemi 3.15 põhjal on tinglik jaotus $X_{n+1}|X_1 = x_1, \ldots, X_n = x_n$ parajasti $\Pi\Big(\alpha + \sum\limits_{i=1}^{n} \delta_{x_i}\Big)$. Järelikult 3.1 põhjal

$$P(X_{n+1} \in A | X_1 = x_1, \ldots, X_n = x_n) = \overline{\Big(\alpha + \sum_{i=1}^{n} \delta_{x_i}\Big)}(A) =$$
$$= \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{M+n} = \frac{M}{M+n}\bar{\alpha}(A) + \frac{1}{M+n}\sum_{i=1}^{n} \delta_{x_i}(A).$$

Niisiis, näeme, et iga n korral vektorid $(X_1, \ldots, X_n, X_{n+1})$ ja $(Y_1, \ldots, Y_n, Y_{n+1})$ on sama jaotusega. $\square$

3.2.1 Hiina restorani protsess

Üks levinumaid viise eelnevas teoreemis kirjeldatud skeemi illustreerimiseks on nn Hiina restorani protsess. Olgu meil esialgu tühi Hiina restoran loenduva arvu laudadega, kusjuures iga laua taha saab istuda loenduv arv kliente. Esimesena saabuv külaline valib suvalise laua. Järgmine külaline istub esimese kõrvale tõenäosusega $\frac{1}{M+1}$ või siis valib uue laua tõenäosusega $\frac{M}{M+1}$. Kui külaline number $n+1$ saabub resorani, siis on seal parajasti n külalist, kes istuvad $k \leq n$ laua taga, kusjuures gruppide suurused on vastavalt $n_1, \ldots, n_k$. Uus külaline valib laua $j \in \{1, \ldots, k\}$ tõenäosusega $\frac{n_j}{M+n}$ või siis istub tühja laua taha tõenäosusega $\frac{M}{M+n}$.
Olgu $P \sim \Pi(\alpha)$ ja $X_1, \ldots, X_n | P = p \overset{\text{i.i.d}}{\sim} p$. Eeldame, et α on mitteatomaarne mõõt. Külalise nr $n+1$ teistega koos laua taha istumine vastab siis varasemalt Dirichlet' protsessist tehtud vaatluse tulemuse kordumisele ehk asjaolule $X_{n+1} \sim \delta_{X_j}$. Uue laua valimine vastab aga asjaolule $X_{n+1} \sim \bar{\alpha}$. Nõndaviisi loob Hiina restorani protsess lihtsa algoritmi Dirichlet' protsessist valimi genereerimiseks. K laua ümber grupeerumine vastab aga sisuliselt indeksite hulga $\{1, \ldots, n\}$ tükeldusele $S^n = \{S_1, \ldots, S_k\}$, mille tekitavad vaatlused Dirichlet' protsessist $X_1, \ldots, X_n$. Indeksid $p, q \in \{1, \ldots, n\}$ kuuluvad hulka S_j, kui $X_p = X_q$.

3.3 Sethuramani tokimurdmine

Eelmises alajaotuses esitasime tokimurdmise algoritmi lõplikumõõtmelise Dirichlet' jaotusega juhusliku vektori genereerimiseks. Väga sarnase põhimõtte järgi on võimalik genereerida ka $\Pi(\alpha)$-jaotusega juhusliku tõenäosusmõõdu ruumil $(S, \mathcal{B}(S))$. Vastavat tõenäosusmõõdu esitusviisi nimetatakse Sethuramani tokimurdmise esituseks ja see leiab laialdast kasutust statistikas. Nimelt võimaldab tokimurdmise protseduur genereerida Dirichlet' protsessi jaotusega tõenäosusmõõtu ligikaudselt. See on eriti oluline rakendustes, kus analüütiline esitus osutub liiga keerukaks. Järgnevalt kirjaldame Sethuramani tokimurdmise algoritmi. Antud alajaotus tugineb [4, lk 41–46].

- Olgu X_1 juhuslik suurus jaotusega $\bar{\alpha}$ ja $V_1 \sim \mathrm{Be}(1, M)$. Tähistame $W_1 = V_1$ ja omistame aatomile X_1 tõenäosusmassi W_1. S.t $P(\{X_1\}) = W_1$ ehk genereeritav tõenäosusmõõt omandab kuju $P = W_1\delta_{X_1}$, kus δ_{X_1} on Diraci mõõt. Intuitiivselt sarnaneb see ühikpikkusega tokist V_1-osa äramurdmisega.

- Olgu X_2 juhuslikust suurusest X_1 sõltumatu juhuslik suurus jaotusega $\bar{\alpha}$ ja $V_2 \sim \text{Be}(1, M)$. Tähistame $W_2 = (1 - V_1)V_2$ ja omistame aatomile X_2 tõenäosusmassi W_2. Sisuliselt murrame esimese sammu järel alles jäänud tokist ära V_2 osa ja seostame saadud "tüki"punktiga X_1. See tähendab, et $P(\{X_2\}) = W_2$ ja genereeritav mõõt omandab kuju $P = W_1\delta_{X_1} + W_2\delta_{X_2}$.

- Olgu X_i juhuslikest suurustest $X_1, \ldots, X_{i-1}$ sõltumatu juhuslik suurus jaotusega $\bar{\alpha}$ ja $V_i \sim \text{Be}(1, M)$. Tähistame $W_i = \Big[\prod_{j=1}^{i-1}(1 - V_j)\Big]V_i$ ja omistame aatomile X_i tõenäosusmassi W_i. Sisuliselt on $\prod_{j=1}^{i-1}(1 - V_j)$ esimese $i - 1$ murdmise järel ühikpikkusega tokist järele jäänud tükk. Murrame sellest vastavlt V_i osa ja seostame selle punktiga X_i. See tähendab, et $P(\{X_2\}) = W_i$ ja genereeritav mõõt omandab kuju $P = \sum_{j=1}^{i} W_j\delta_{X_j}$.

Jätkates seda protsessi lõpmatuseni, osutub, et saadud juhuslik tõenäosusmõõt $P = \sum_{i=1}^{\infty} W_i\delta_{X_i}$ on parajasti $\Pi(\alpha)$-jaotusega.
On selge, et iga tokimurdmise algoritmi teel genereeritud tõenäosusmõõt on diskreetne. Üldisemalt saab näidata, et Dirichlet' protsessi fikseeritud väärtus on peaaegu kindlasti diskreetne. Selle asjaolu valguses tuleme tagasi Dirichlet' protsessist tehtud vaatluste mudeli juurde. Valides jaotuse $\Pi(\alpha)$ juhusliku tõenäosusmõõdu P eeljaotuseks välistame formaalselt võimaluse, et vaatlused Dirichlet' protsessist $X_1, \ldots, X_n$ on mõne pideva jaotusega (nt normaaljaotusega) juhuslikud suurused. Praktikas aga ei ole Dirichlet' protsessi diskreetsus üldjuhul suur probleem, sest saame pidevaid jaotusi kuitahes täpselt diskreetsete jaotustega lähendada. Viimane asjaolu haakub teoreemiga 0.7, kus tõestasime, et teatud diskreetsete jaotuste alamhulk on kõikjal tihe ruumis $\mathcal{P}$.
Ka Dirichlet' protsessi jaotuse $\Pi(\alpha)$ kandja – vähim kinnine hulk, mille mõõt on 1, mida tähistame kui $\text{supp}(\Pi(\alpha))$ – saab sobiva baasimõõdu α korral olla kogu ruum $\mathcal{P}$, s.t sisaldada ka kõiki pidevaid jaotusi. Üldisemalt kehtib võrdus

$$\text{supp}(\Pi(\alpha)) = \{p \in \mathcal{P} : \text{supp}(p) \subset \text{supp}(\bar{\alpha})\}.$$

Märgime, et juhul, kui $\bar{\alpha}$ kuulub näiteks normaaljaotuste perre, siis $\text{supp}(\bar{\alpha}) = S$ ja seega iga $p \in \mathcal{P}$ korral $\text{supp}(p) \subset \text{supp}(\bar{\alpha})$ ning vastava baasimõõdu $\alpha = M\bar{\alpha}$ on jaotuse $\Pi(\alpha)$ kandjaks kogu tõenäosusmõõtude ruum $\mathcal{P}$.

Kirjandus

[1] P. Billingsley, *Convergence of Probability Measures.* John Wiley & Sons, New York, 1999.

[2] V. Bogachev, *Weak Convergence of Measures.* American Mathematical Society, Providence, 2018.

[3] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, *The Annals of Statistics,* **1**, 209–230, 1973.

[4] S. Ghosal, *The Dirichlet' process, related priors and posterior asymptotics.* Ilmunud: N. L. Hjort, C. Holmes, P. Müller, S. G. Walker, toim. *Bayesian Nonparametrics.* Cambridge University Press, Cambridge, 2010.

[5] S. Ghosal, A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press, Cambridge, 2017.

[6] J. Lember, *Tõenäosusteooria II* (loengukonspekt). Tartu Ülikool, 2018.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Mikael Raihhelgauz (sünnikuupäev 13.04.1999),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

   "Dirichlet' protsess",

   mille juhendaja on Jüri Lember,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Tartu, 13.05.2020