Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Jaan Erik Pihel

Kolmekaupa Markovi ahelate Viterbi raja lähendamine variatsiooniliste meetoditega

Matemaatika ja statistika õppekava

Matemaatika eriala

Magistritöö (30 EAP)

Juhendajad: Prof. Jüri Lember
MSc. Oskar Soop

Tartu 2025

KOLMEKAUPA MARKOVI AHELATE VITERBI RAJA LÄHENDAMINE VARIATSIOONILISTE MEETODITEGA

Magistritöö

Jaan Erik Pihel

Lühikokkuvõte

Kolmekaupa Markovi ahel (TMM) üldistab paarikaupa Markovi ning varjatud Markovi ahelaid. Ülesande konstrueerimisel eeldame, et meil on Markovi protsessi (U, X, Y) puhul fikseeritud vaatlusandmed $(\boldsymbol{y}_t)_{t=1}^T$ ning me soovime leida Viterbi rada $\boldsymbol{x}^*$ ehk $\arg\max_{\boldsymbol{x}} \sum_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{x}|\boldsymbol{y})$. Et selle lahendamine on NP raske, leitakse töös Viterbi raja lähend variatsioonise Bayesi meetodi abil. *Belief propagation* (BP) ja *variational message passing* (VMP) algoritmide tulemusena leitakse erinevate kitsendustega $q(\boldsymbol{u}, \boldsymbol{x})$, mis minimiseerib KL kaugust $D[q\|p]$ ning seejärel antakse lähend mõõdu $q(\boldsymbol{x})$ Viterbi rajana kasutades Viterbi algoritmi. Eksperimendid näitavad, et kaks algoritmi komplementeerivad üksteist ehk arvutuslikult keerukam BP algoritm ei ole alati parem.

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

Märksõnad: HMM, PMM, TMM, juhuslikud protsessid, Markovi ahelad, statistiline järeldamine.

APPROXIMATING VITERBI PATH IN TRIPLET MARKOV CHAINS USING VARIATIONAL METHODS

Master thesis

Jaan Erik Pihel

Abstract

Triplet Markov chain (TMM) generalises both pairwise Markov and hidden Markov chains. We look at a problem for which we have a Markov process (U, X, Y) with fixed observations $(\boldsymbol{y}_t)_{t=1}^T$ and we wish to find a Viterbi path for $\boldsymbol{x}^*$, i.e. $\arg\max_{\boldsymbol{x}} \sum_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{x}|\boldsymbol{y})$. Since that is a NP hard problem, we choose to approximate the Viterbi path using variational Bayes methods. Belief propagation (BP) and variational message passing (BP) algorithms find a measure $q(\boldsymbol{u}, \boldsymbol{x})$ under different constraints which minimise KL divergence $D_{\mathrm{KL}}[q\|p]$ and give the Viterbi path approximation as the Viterbi path of the marginal $q(\boldsymbol{x})$ using the Viterbi algorithm. Experiments show that the two algorithms complement each other, i.e. sometimes the computationally easier VMP algorithm outperforms the other algorithm.

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

Key Words: HMM, PMM, TMM, random processes, Markov chains, statistical inference.

Contents

Sissejuhatus

Diskreetsed Markovi ahelad ja varjatud Markovi ahelad ning nende omadused on varasemalt põhjalikult uuritud. Kui varjatud Markovi ahela $\{W_t, Y_t\}_{t=1}^T$ puhul sõltub vaatlus Y_t vaid juhuslikust suurusest W_t, vaatleme me laiendatud mudeleid mida kutsutakse paarikaupa Markovi mudeliteks (PMM), kus $p(y_t|w_{t-1}, y_{t-1}, w_t) \neq p(y_t|w_t)$. Kui nõuda, et protsess $\{W_t\} = \{U_t, X_t\}$ oleks omakorda paarikaupa Markovi ahel, saame kolmekaupa Markovi mudeli (TMM) $\{U_t, X_t, Y_t\}_{t=1}^T$. Kui vaatlusandmed $\boldsymbol{y}$ on fikseeritud, saame huvipakkuva diskreetse mittehomogeense paarikaupa Markovi mudeli kirjeldamaks ahela $\{U_t, X_t|\boldsymbol{y}\}$ jaotust.

Kui varjatud Markovi ahela puhul on Viterbi raja ehk kõige suurema tõenäosusega raja $\boldsymbol{x}$ leidmine lihtne Viterbi algoritmi abil, siis on näidatud, et juba PMM korral võib Viterbi raja leidmine olla NP raske. Kuigi *variational inference* meetodid leiavad kirjanduses enamasti rakendust jaotuste lähendamisel, loome magistritöös kaks algoritmi, mis annavad punkthinnanguna Viterbi algoritmi lähendi kasutades variatsioonilisi meetodeid. Kirjeldame, kuidas leida teatud kitsendustega jaotused, mis on Kullback-Leibleri kauguse mõttes lähedaseimad huvipakkuvale paarikaupa Markovi mudelile ning seejärel saame leitud jaotustele rakendada Viterbi algoritmi. Sellele eelnevalt tõestame, et kõige lähedasemad jaotused eksisteerivad ning algoritmid nendeks ka koonduvad.

Eksperimentaalosas tegeleme diskreetsete mittehomogeensete paarikaupa Markovi ahelate kirjeldustega, mille põhjal indikeerida, kumba töös kirjeldatud algoritmi rakendada. Samuti mõõdame, milline algoritm keskeltläbi paremaid tulemusi annab.

1 Variatsioonilised Bayesi meetodid Viterbi raja lähendamiseks

1.1 Paarikaupa ja kolmekaupa Markovi ahelad

Et paarikaupa ja kolmekaupa Markovi ahelate omadusi on varasemalt uuritud palju (Kuljus **and** Lember, 2022), (Soop, 2023), (Avans, K., 2021), siis toome välja vaid olulisemad definitsioonid ja tulemused.

Meenutame, et diskreetne Markovi ahel $\boldsymbol{Z} = (Z_t)_t$ on protsess mille korral kehtib Markovi omadus ehk iga $T \in \mathbb{N}$ korral

$$P(Z_1 = z_1, \ldots, Z_T = z_T) = P(Z_1 = z_1) \prod_{t=2}^{T} P(Z_t = z_t | Z_{t-1} = z_{t-1}),$$

kus Z_t võtab väärtusi ülimalt loenduval hulgal $\mathcal{Z}$. Kui iga t korral on üleminekumaatriks $P(Z_t = z_t | Z_{t-1} = z_{t-1})$ sama, öeldakse, et see Markovi ahel on homogeenne, vastasel juhul mittehomogeenne.

Paarikaupa Markovi ahelaks nimetame kahe muutujaga Markovi ahelat $\{Z_t\}_t = \{(X_t, Y_t)\}_t$, mille võimalikud väärtused on hulgas $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$. Olgu teada juhusliku suuruse Z_1 tihedus π korrutismõõdu $c \times \mu$ suhtes, kus c on loendav mõõt tõenäosusruumiga $(\mathcal{X}, \mathcal{B}(\mathcal{X}), c)$ ning μ mõõt tõenäosusruumiga $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mu)$. Kui $\mathcal{Y}$ on ülimalt loenduv, saame Markovi ahelat kirjeldada $t > 1$ korral üleminekumaatriksite $P(Z_t = z_t | Z_{t-1} = z_{t-1})$ abil. Kui $\mathcal{Y}$ ei ole ülimalt loenduv, kirjeldame Markovi ahelat üleminekutiheduse abil

$$p : \mathcal{Z}^2 \to [0, \infty), \ (z_t, z_{t-1}) \mapsto p_t(z_t | z_{t-1}),$$

kus $p_t(\cdot | z_{t-1})$ on tõenäosusmõõdu tihedusfunktsioon korrutismõõdu $c \times \mu$ suhtes. Saab

aga näidata, et fikseeritud $\boldsymbol{y} \in \mathcal{Y}^T$ korral tingliku protsessi $\boldsymbol{X}|\boldsymbol{y}$ tihedus

$$\begin{aligned} P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{y}) &= P(X_1 = x_1|\boldsymbol{Y} = \boldsymbol{y}) \prod_{t=2}^{T} P(X_t = x_t|X_{t-1} = x_{t-1}, \boldsymbol{Y} = \boldsymbol{y}) \\ &= P(X_1 = x_1|\boldsymbol{Y} = \boldsymbol{y}) \prod_{t=2}^{T} \frac{P(X_t = x_t, X_{t-1} = x_{t-1}|\boldsymbol{Y} = \boldsymbol{y})}{P(X_{t-1} = x_{t-1}|\boldsymbol{Y} = \boldsymbol{y})} \end{aligned}$$

diskreetne (Avans, K., 2021 lk 15). Seega on meid edaspidi huvitavad jaotused igal juhul diskreetsed.

Olgu teada homogeense kahekordse Markovi ahela üleminekutihedus $p(x_t, y_t|x_{t-1}, y_{t-1})$ ning algtihedus $\pi(x_1, y_1)$. Näitame, et fikseeritud $\boldsymbol{y} \in \mathcal{Y}^T$ korral on tinglik protsess $\boldsymbol{X}|\boldsymbol{y}$ Markovi omadusega tõenäosusmõõt ja faktoriseerub kujul

$$p(\boldsymbol{x}|\boldsymbol{y}) = \pi(x_1|\boldsymbol{y}) \prod_{t=2}^{T} p(x_t|x_{t-1}, \boldsymbol{y}_{t-1:T}). \tag{1.1}$$

Tähistame fikseeritud $i < j$ korral vektoreid kui $\boldsymbol{x}_{i:j} := (x_i, \ldots, x_j)$, $\boldsymbol{x} := (x_t)_{t=1}^T$ ning $\boldsymbol{x}_{\setminus i,j} := (x_1, \ldots, x_{i-1}, x_{i+1}, ots, x_{j-1}, x_{j+1}, \ldots, x_T)$. Kirjutame

$$\begin{aligned} p(x_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{y}) &= \frac{p(x_t, \boldsymbol{y}_{t+1:T}|\boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1})\ p(\boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_{t+1:T}, y_t|\boldsymbol{y}_{1:t-1}, \boldsymbol{x}_{1:t-1})\ p(\boldsymbol{x}_{1:t-1}, \boldsymbol{y}_{1:t-1})} \\ &= \frac{p(x_t, \boldsymbol{y}_{t:T}|x_{t-1}, y_{t-1})}{p(\boldsymbol{y}_{t:T}|x_{t-1}, y_{t-1})} \\ &= p(x_t|x_{t-1}, \boldsymbol{y}_{t-1:T}), \end{aligned}$$

kus esimeses võrduses kasutame Bayesi valemit ning teises kasutame ära seda, et tegemist on varjatud Markovi ahelaga, ning kolmandas taas Bayesi valemit.

Analoogse arutelu tulemusel saame ka, et

$$p(x_t|x_{t-1}, \boldsymbol{y}_{t-1:T}) = \frac{p(x_t, \boldsymbol{y}_{t:T}|x_{t-1}, y_{t-1})}{p(\boldsymbol{y}_{t:T}|x_{t-1}, y_{t-1})} = \frac{p(x_t, \boldsymbol{y}_{t:T}|x_{t-1}, \boldsymbol{y})}{p(\boldsymbol{y}_{t:T}|x_{t-1}, \boldsymbol{y})} = p(x_t|x_{t-1}, \boldsymbol{y}).$$

Edaspidi saame sõltuvust suurusest $\boldsymbol{y}$ vaadelda implitsiitselt ning defineerida tõenäo-

susmõõdu $p(\boldsymbol{x})$, mis faktoriseerub kujul (1.1) ehk $\boldsymbol{X}$ on diskreetne mittehomogeenne Markovi ahel.

Nüüd defineerime kolmekaupa Markovi ahela kui $\{Z_t\}_t = \{(U_t, X_t, Y_t)\}_t$ üle ruumi $\mathcal{Z} \subseteq \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$, kus lisaks varasemale on defineeritud lõplik $\mathcal{U} = \{1, \ldots, |\mathcal{U}|\}$ loendava mõõduga c'. Algtihedus olgu $\pi(u_1, x_1, y_1)$ ning üleminekutihedus

$$p : \mathcal{Z}^2 \to [0, \infty),\ (z_t, z_{t-1}) \mapsto p(z_t | z_{t-1}).$$

Kuna paarikaupa ja kolmekaupa Markovi ahelad on Markovi ahelad, omavad nad omaette definitsioonidena mõtet, kui anda marginaalprotsessidele $\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Y}$ asjakohane tähendus. Juhuslik vektor $\boldsymbol{X} \in \mathcal{X}^T$ kirjeldab meile huvipakkuvaid varjatud tunnuseid, juhuslik vektor $\boldsymbol{U} \in \mathcal{U}^T$ olgu segav parameeter ning juhuslik vektor $\boldsymbol{Y} \in \mathcal{Y}^T$ kirjeldagu vaatlusandmeid. Ehk me tähistame vektoriga $\boldsymbol{y}$ juhusliku vektori $\boldsymbol{Y}$ realisatsiooni, mille järgi tinglikustades saame diskreetse mittehomogeense paarikaupa Markovi ahela $(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{y}$ ning juhusliku vektori $\boldsymbol{U}$ ühisjaotusest välja summeerides saame juhusliku vektori $\boldsymbol{X}|\boldsymbol{y}$ jaotuse.

Et iga kolmekaupa Markovi ahel on ka paarikaupa Markovi ahel, siis nagu näitasime paarikaupa Markovi ahela puhul, saame ka siin vaadelda tingliku protsessi $(\boldsymbol{U}, \boldsymbol{X})|\boldsymbol{y}$ asemel sellele vastavat mittehomogeenset paarikaupa Markovi ahelat. Kirjutame edaspidi sellise diskreetse mittehomogeense paarikaupa Markovi ahela algjaotuse kui π ning iga $t > 2$ korral üleminekumaatriksi

$$p_t : (\mathcal{U} \times \mathcal{X})^2 \to [0, \infty),\ (u_{t-1}, x_{t-1}, u_t, x_t) \mapsto p_t(u_t, x_t | u_{t-1}, x_{t-1}).$$

Uuritav mudel on $\boldsymbol{u} \in \mathcal{U}^T, \boldsymbol{x} \in \mathcal{X}^T$ jaoks

$$p(\boldsymbol{u}, \boldsymbol{x}) = \pi(u_1, x_1) \prod_{t=2}^{T} p_t(u_t, x_t | u_{t-1}, x_{t-1}). \tag{1.2}$$

Marginaalmõõte tähistame kui $p_x(\boldsymbol{x}) := \sum_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{x})$ ja $p_u(\boldsymbol{u}) := \sum_{\boldsymbol{x}} p(\boldsymbol{u}, \boldsymbol{x})$, kusjuures

vastavad marginaalprotessid pole üldiselt Markovi ahelad (Soop, 2023).

1.1.1 Näited

Kirjeldame vahelduva režiimiga mudeleid nagu seda tehti peatükis 3.2.3 (Avans, K., 2021). Olgu meil Markovi protsess $\{U_t\}$, kus U_t võtab väärtusi hulgal $\{A, B, C\}$, mida nimetatakse režiimideks. Olgu Markovi protsessi kirjeldav algtõenäosus π_u ning 3×3 üleminekumaatriks p_u. Tavaks on nõuda, et see üleminekumaatriks on ühikmaatriksile lähedane, et järgmisel ajasammul samasse režiimi jäämine oleks suure tõenäosusega sündmus ehk $p_u(u_t = a|u_{t-1} = a) = 1 - \varepsilon_a$ iga $a \in \{1, 2, 3\}$ korral. Ütleme, et ajahetkel t on protsess režiimis A kui $u_{t-1} = A, u_t = A$.

Vaatleme lisaks Markovi protsessile $\{U_t\}$ veel lisaks kaheseisundilist protsessi $\{X_t\}, X_t \in \{0, 1\}$. Olgu igas režiimis X_t jaotus kirjeldatav Markovi protsessina - st ajahetkel t olgu režiimis A protsessi X_t üleminekumaatriks 2×2 maatriks p_A. Igale režiimile loome ka vastava algtõenäosustejaotuse $p(x_1|u_1)$. On loomulik nõuda, et p_A, p_B, p_C erineksid üksteisest piisavalt.

Nüüd saame kirjeldada protsessi $\{U_t, X_t\}$ **kahekaupa Markovi ahela** abil, mille jaotust saab kirjeldada algtõenäosuse ja üleminekutõenäosusega

$$\pi(u_1, x_1) = \pi_u(u_1)p(x_1|u_1)$$
$$p(u_t, x_t|u_{t-1}, x_{t-1}) = p_u(u_t|u_{t-1})p(x_t|x_{t-1}, u_{t-1}, u_t),$$

kus on veel $p(x_t|x_{t-1}, u_{t-1}, u_t)$ on iga $u_t \neq u_{t-1}$ korral mingi 2×2 üleminekumaatriks p_{u_{t-1},u_t}. Saame kahekordse Markovi ahela üleminekumaatriksi

$$p = \begin{pmatrix} p_u(A|A)p_A & p_u(A|B)p_{AB} & p_U(A|C)p_{AC} \\ p_u(B|A)p_{BA} & p_u(B|B)p_B & p_U(B|C)p_{BC} \\ p_u(C|A)p_{CA} & p_u(C|B)p_{CB} & p_U(C|C)p_{CC} \end{pmatrix}.$$

Ütleme, et juhuslikku suurust X_t loetakse müraga, kui meil on vaatlusandmeid kirjeldav

juhuslik suurus Y_t, mille jaotus on $p(y_t|x_t) = x_t + \varepsilon$, kus ε on mingi sõltumatu juhuslik suurus, nt $\varepsilon \sim \mathcal{N}(0,1)$. Saame nüüd protsessi $\{U_t, X_t, Y_t\}$ kirjeldada **kolmekaupa Markovi ahelana**, kus

$$\pi(u_1, x_1, y_1) = \pi_u(u_1)p(x_1|u_1)p(y_1|x_1)$$
$$p(u_t, x_t, y_t|u_{t-1}, x_{t-1}, y_{t-1}) = p_u(u_t|u_{t-1})p(x_t|x_{t-1}, u_{t-1}, u_t)p(y_t|x_t).$$

Kui vaatlusandmed $\boldsymbol{y}$ on fikseeritud, saame protsessi $\{U_t, X_t\}|\boldsymbol{y}$ kirjeldada nüüd **mittehomogeense paarikaupa Markovi ahelana**, kus algtõenäosus ja üleminekumaatriksid on

$$\pi(u_1, x_1) = \frac{\sum_{\boldsymbol{u}_{2:T}, \boldsymbol{x}_{2:T}} p(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})}$$
$$p_t(u_t, x_t|u_{t-1}, x_{t-1}) = \frac{\sum_{\boldsymbol{u}_{1:t-1}, \boldsymbol{x}_{1:t-1}} \sum_{\boldsymbol{u}_{t+1:T}, \boldsymbol{x}_{t+1:T}} p(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})},$$

kus $p(\boldsymbol{y}) = \sum_{\boldsymbol{u}, \boldsymbol{x}} p(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y})$.

1.2 Ülesande ja lähendite kirjeldus

Olgu meil lõplikul olekute ruumil defineeritud mittehomogeene paarikaupa Markovi ahel $(\boldsymbol{X}, \boldsymbol{U})$. Ülesandeks on leida $\boldsymbol{x}^* \in \mathcal{X}^T$, kus

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in \mathcal{X}^T} \sum_{\boldsymbol{u} \in \mathcal{U}^T} p(\boldsymbol{x}, \boldsymbol{u}). \tag{1.3}$$

Sellist $\mathbf{x}^*$ nimetatakse **Viterbi rajaks**. On võimalik näidata (Lyngsø **and** Pedersen, 2002), et Viterbi raja leidmine on NP-raske ülesanne. Viterbi raja täpselt leidmise asemel leiame marginaalmõõdule p_x lähendi q_x, mille jaoks uurime, kas mingitel juhtudel on lähendi q_x Viterbi rada sagedasti võrdne Viterbi rajaga $\boldsymbol{x}^*$. Neid tingimusi uurime eksperimentide abil peatükis 4. Selles töös kasutame marginaalmõõdu lähendamiseks variatsioonilisi meetodeid. Töö koosneb kahe algoritmi kirjeldamisest, mis mõlemad

lahendavad erinevate kitsendustega minimiseerimisülesannet

$$\arg\min_{q} D_{\mathrm{KL}}[q\|\ p], \tag{1.4}$$

kus Kullback-Leibleri kaugust defineeritakse kui

$$D_{\mathrm{KL}}[q\|\ p] := \sum_{\boldsymbol{x}\in\mathcal{X}^T,\boldsymbol{u}\in\mathcal{U}^T} q(\boldsymbol{u},\boldsymbol{x}) \ln \frac{q(\boldsymbol{u},\boldsymbol{x})}{p(\boldsymbol{u},\boldsymbol{x})}. \tag{1.5}$$

1. Esimeses *belief propagation* (BP) algoritmis leiame KL kauguse (1.5) mõttes parima tõenäosusmõõdu kujul $q(\boldsymbol{u},\boldsymbol{x}) = q_x(\boldsymbol{x})q_u(\boldsymbol{u})$ ning seejärel kasutame Viterbi algoritmi mõõdu q_x jaoks, et saada Viterbi raja lähend $\boldsymbol{x}^*_{\mathrm{BP}}$.

2. Teises *variational message passing* (VMP) algoritmis nõuame ajalist sõltumatust tõenäosusmõõdus $q(\boldsymbol{u},\boldsymbol{x}) = \prod_{t=1}^{T} q_t(u_t,x_t)$ ning minimiseerime KL kaugust (1.5). Viterbi raja lähendi leiame $x_t^* = \arg\max_{x_t} \sum_{u_t} q_t(u_t,x_t)$ abil iga t korral saades Viterib raja lähendi $\boldsymbol{x}^*_{\mathrm{VMP}}$.

Algoritmide nimed on inspireeritud artiklist (Parr **andothers**, 2019).

Peatükis 1.4 uurime põgusalt, miks me minimiseerimisülesande (1.4) asemel ei vaata ülesannet

$$\arg\min_{q} D_{\mathrm{KL}}[p\|\ q],$$

kus KL kauguse sees on mõõdud p ja q vahetanud oma asukohad.

1.3 Variatsiooniline meetod mõõdu parandamiseks

Fikseerime $T \in \mathbb{N}$. Olgu $\mathcal{X}$ ja $\mathcal{U}$ lõplikud tähestikud ja $p(\boldsymbol{u},\boldsymbol{x})$ olgu mingi hulgal $\mathcal{X}^T \times \mathcal{U}^T$ antud diskreetne tõenäosusmõõt. Olgu $\mathcal{P}(\mathcal{U})$ ja $\mathcal{P}(\mathcal{X})$ vastavalt hulgal $\mathcal{U}^T$ ja $\mathcal{X}^T$ antud kõikide diskreetsete tõenäosusmõõtude hulk. Suvalise paari $q_x \in \mathcal{P}(\mathcal{X})$ ja $q_u \in \mathcal{P}(\mathcal{U})$ korral olgu $q_u \times q_x$ korrutismõõt hulgal $\mathcal{U}^T \times \mathcal{X}^T$. Vaatleme optimiseerimisülesannet

$$\min_{q_u\times q_x} D_{\mathrm{KL}}[q_u \times q_x \| p], \tag{1.6}$$

kus miinimum võetakse üle kõikide korrutismõõtude.

Näitame, et see miinimum leidub. Selleks piisab meenutada, et diskreetsete tõenäosusmõõtude kaalud moodustavad $|\mathcal{U}|^T-1$ ja $|\mathcal{X}|^T-1$ dimensionaalsed simpleksid, mis on kompaktsed ruumid ja mille otsekorrutis on samuti kompaktne ruum. KL kaugus on esimese argumendi järgi pidev ning on mittenegatiivne. Funktsionaalanalüüsist on teada Weierstrassi teoreem, (E. Oja **and** P. Oja, 1991 II peatükk §5) mille põhjal miinimum ka saavutatakse.

Järgmiseks uurime, millises $\mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{X})$ alamruumis $\mathcal{P}_u \times \mathcal{P}_x$ on KL kaugus esimese argumendi, st $q_u \times q_x$, järgi rangelt kumer. KL kaugus (1.5) on lõplik, kui $\operatorname{supp} q_u \times q_x \subseteq \operatorname{supp} p$, sest uuritavas KL kauguses on liidetavaid lõplik hulk ning liidetav ei ole lõplik vaid juhul, kui leiduvad sellised $\boldsymbol{u}, \boldsymbol{x}$ nii, et $q_u(\boldsymbol{u})q_x(\boldsymbol{x}) > 0$ ja $p(\boldsymbol{u}, \boldsymbol{x}) = 0$. Olgu S_u, S_x sellised simpleksid, kus iga $q_u \in S_u, q_x \in S_x$ korral $\operatorname{supp} q_u \times q_x \subseteq \operatorname{supp} p$. Mittetriviaalselt eeldame, et selliste simpleksite korrutis ei ole tühihulk. Nüüd näitame, et KL kaugus on rangelt kumer ehk iga $q_u \times q_x, \tilde{q}_u \times \tilde{q}_x$ korral, kus $q_u \times q_x \neq \tilde{q}_u \times \tilde{q}_x$ ja $D_{\mathrm{KL}}[q_u \times q_x \| p]$, $D_{\mathrm{KL}}[\tilde{q}_u \times \tilde{q}_x \| p]$ on lõplikud kehtib

$$D_{\mathrm{KL}}[\lambda(q_u \times q_x) + (1-\lambda)(\tilde{q}_u \times \tilde{q}_x)||p] < \lambda D_{\mathrm{KL}}[q_u \times q_x||p] + (1-\lambda)D_{\mathrm{KL}}[\tilde{q}_u \times \tilde{q}_x||p].$$

Mittenegatiivsete arvude $a_1, \ldots, a_n, a := \sum_{k=1}^n a_k, b_1, \ldots, b_n, b := \sum_{k=1}^n b_k$ korral kehtib logaritmide summa võrratus (Log sum inequality, 2025)

$$a \ln \frac{a}{b} \leq \sum_{k=1}^{n} a_k \ln \frac{a_k}{b_k},$$

kus võrdus kehtib vaid juhul, kui leidub c nii, et $a_k \equiv cb_k$. Saame soovitud kuju kui võrratuse paremal pool logaritmide argumendid korrutada liikmega $\frac{\lambda}{\lambda}$ või $\frac{1-\lambda}{1-\lambda}$ ja kirjutada

iga $\boldsymbol{u}, \boldsymbol{x}$ korral

$$(\lambda q_u(\boldsymbol{u})q_x(\boldsymbol{x}) + (1-\lambda)\tilde{q}_u(\boldsymbol{u})\tilde{q}_x(\boldsymbol{x})) \ln \frac{\lambda q_u(\boldsymbol{u})q_x(\boldsymbol{x}) + (1-\lambda)\tilde{q}_u(\boldsymbol{u})\tilde{q}_x(\boldsymbol{x})}{\lambda P(\boldsymbol{u},\boldsymbol{x}) + (1-\lambda)P(\boldsymbol{u},\boldsymbol{x})}$$
$$\leq \lambda q_u(\boldsymbol{u})q_x(\boldsymbol{x}) \ln \frac{\lambda q_u(\boldsymbol{u})q_x(\boldsymbol{x})}{\lambda P(\boldsymbol{u},\boldsymbol{x})} + (1-\lambda)\tilde{q}_u(\boldsymbol{u})\tilde{q}_x(\boldsymbol{x}) \ln \frac{(1-\lambda)\tilde{q}_u(\boldsymbol{u})\tilde{q}_x(\boldsymbol{x})}{(1-\lambda)P(\boldsymbol{u},\boldsymbol{x})},$$

kusjuures et $q_u \times q_x \neq \tilde{q}_u \times \tilde{q}_x$, siis leiduvad $\boldsymbol{u}, \boldsymbol{x}$ nii, et eelmine võrratus on range.

Oleme näidanud, et ülesande (1.6) lahend leidub ja on ühene. Järgmiseks uurime, kuidas lahendit leida.

Esiteks uurime, kas mõõdu p marginaaljaotuste korrutis $p_u \times p_x$ on ülesande (1.6) lahendiks. Toome kontranäite, et see nii ei ole. Vaatleme ühisjaotust $p = [0.1, 0.2; 0.1, 0.6]$ (st $p(u = 1, x = 2) = 0.2$), marginaale $p_u = [0.3, 0.7], p_x = [0.2, 0.8]$ ja mõõte $q_u = p_u$, $q_x = [0.19, 0.81]$. Leides KL kaugused näeme, et $D_{\mathrm{KL}}[P_u \times P_x||P] > D_{\mathrm{KL}}[Q_u \times Q_x||P]$. Kuigi leitud lähendid ei ole üldiselt marginaalid, võib selline lähenemine olla Viterbi raja leidmisel ikkagi hea, kuid selle hindamiseks kasutame eksperimente.

Uurime miks on ülesande (1.6) lahendite q_u, q_x analüütiliselt leidmine raske. Proovime otsida Lagrange'i määramata kordajate meetodi abil sõna $\tilde{\boldsymbol{u}} \in \mathcal{U}^T$ kaalu $q_u(\tilde{\boldsymbol{u}})$ jaoks kriitilist kohta

$$\frac{\partial}{\partial q_u(\tilde{\boldsymbol{u}})} \left[D_{\mathrm{KL}}[q_u \times q_x \| p] + \lambda_1 \Big(1 - \sum_x q_x\Big) + \lambda_2 \Big(1 - \sum_u q_u\Big) \right] = 0,$$

siis leides tuletised saame võrrandi

$$1 + \ln q_u(\tilde{\boldsymbol{u}}) - \sum_{\boldsymbol{x}} q_x(\boldsymbol{x}) \ln p(\tilde{\boldsymbol{u}}, \boldsymbol{x}) + \lambda_2 = 0,$$

mis on meile probleemiks, sest $q_x(\boldsymbol{x})$ kaalud leiaksime analoogselt kasutades mõõtu q_u sh kaalu $q_u(\tilde{\boldsymbol{u}})$, mida me ei pruugi teada.

Et ülesande (1.6) lahendamine analüütiliselt on raske, siis loome iteratiivse algoritmi lähendite q_u, q_x leidmiseks. Me alustame suvaliste tõenäosusmõõtudega $q_u^{(0)}, q_x^{(0)}$. Igal iteratsioonil fikseerime $q_x^{(i)}$ leidmaks uut tõenäosusmõõtu $q_u^{(i+1)}$, mis vähendab KL kau-

gust (1.5), misjärel fikseerime $q_u^{(i+1)}$ ning leiame $q_x^{(i+1)}$, mis samuti KL kaugust vähendab. Kirjeldame, kuidas selline algoritm võiks käituda. Valime algjaotused

$$q_u^{(0)} S_u,\ q_x^{(0)} \in S_x,\ d_0 := D_{\mathrm{KL}}[q_u^{(0)} \times q_x^{(0)} || p] < \infty$$

ning seejärel leiame

$$q_u^{(i+1)} = \underset{q_u \in \mathcal{P}(\mathcal{U})}{\arg\min}\, D_{\mathrm{KL}}[q_u \times q_x^{(i)} || p], \qquad d_{2i-1} := D_{\mathrm{KL}}[q_u^{(i+1)} \times q_x^{(i)} || p] \tag{1.7}$$

$$q_x^{(i+1)} = \underset{q_x \in \mathcal{P}(\mathcal{X})}{\arg\min}\, D_{\mathrm{KL}}[q_u^{(i+1)} \times q_x || p], \qquad d_{2i} := D_{\mathrm{KL}}[q_u^{(i+1)} \times q_x^{(i+1)} || p]. \tag{1.8}$$

Et iga $i > 0$ korral $q_u^{(i)} \in \mathcal{P}(\mathcal{U})$, siis

$$\min_{q_u \in \mathcal{P}(\mathcal{U})} D_{\mathrm{KL}}[q_u \times q_x^{(i)} \| p] \leq D_{\mathrm{KL}}[q_u^{(i)} \times q_x^{(i)} \| p]$$

ja sarnaselt saame arutleda ka q_x puhul. Seepärast on iteratsioonidel leitavad mõõdud sellised, mis moodustavad KL kauguste mõttes monotoonselt mittekasvava jada (d_k). Võime vaadelda iteratsioone kui kiireima languse meetodi (*gradient descent*) erijuhtu. Et KL kaugus on mittenegatiivne, on see jada alt tõkestatud ning koondub. Defineerime

$$L : S_u \times S_x \to \mathbb{R},\ (q_u, q_x) \mapsto D_{\mathrm{KL}}\left[q_u \times q_x \| p\right]$$

Järgmisena näitame, et algoritmi protseduur koondub ülesande 1.6 lahendini. Täpsemalt, et toimub koondumine

$$\left(q_u^{(k)}, q_x^{(k)}\right)_k \xrightarrow{k} \left(q_u^{(\infty)}, q_x^{(\infty)}\right)$$

nii, et

$$L\left(q_u^{(\infty)}, q_x^{(\infty)}\right) = \min_{q_u \in S_u, q_x \in S_x} L(q_u, q_x).$$

Selle sõnastame hiljem teoreemina, kuid enne tõestame vahetulemusena järgnevad lem-

mad.

Lemma 1. *Olgu* $(\tilde{q}_u, \tilde{q}_x)$ *ruumi* $S_u \times S_x$ *element, millele vastav KL kaugus ei ole minimaalne. Siis algoritmi sammudega* (1.7) *või* (1.8) *saadav paar* $(\tilde{q}'_u, \tilde{q}_x)$ *või* $(\tilde{q}_u, \tilde{q}'_x)$ *on väiksema KL kaugusega.*

Tõestus. Olgu $(\tilde{q}_u, \tilde{q}_x)$ selline paar, mille korral

$$L(\tilde{q}_u, \tilde{q}_x) > \min_{q_u \in S_u, q_x \in S_x} L(q_u, q_x).$$

Et KL kaugus on q kaalude järgi diferentseeruv ja rangelt kumer, siis leidub vaid üks kriitiline koht – selle globaalne miinimum. Seepärast $\nabla_{q_u,q_x} L(q_u, q_x) \neq \mathbf{0}$. On lihtne näha, et kui $\nabla_{q_u} L(q_u, q_x) = \mathbf{0}$ ja $\nabla_{q_x} L(q_u, q_x) = \mathbf{0}$, siis ka $\nabla_{q_u,q_x} L(q_u, q_x) = \mathbf{0}$. Seega $\nabla_{q_u} L(\tilde{q}_u, \tilde{q}_x) \neq \mathbf{0}$ või $\nabla_{q_x} L(\tilde{q}_u, \tilde{q}_x) \neq \mathbf{0}$, üldisust kitsendamata eeldame, et $\nabla_{q_u} L(\tilde{q}_u, \tilde{q}_x) \neq \mathbf{0}$ ning KL kauguse diferentseruvusest tulenevalt leidub $\tilde{q}_u$ mingis ümbruses selline q''_u, et kehtib

$$L(\tilde{q}_u, \tilde{q}_x) > L(q''_u, \tilde{q}_x) \geq L(\tilde{q}'_u, \tilde{q}_x),$$

kusjuures teine võrratus põhineb valemi (1.7) minimaalsusel. □

Järeldus 1. *Sammude* (1.7) *ja* (1.8) *järjest rakendamisel on täpselt üks püsipunkt – probleemi 1.6 lahend.*

Paraku püsipunkti olemasolu ja KL kauguse rangelt kahanemine ei garanteeri koonduvust püsipunktini. Tähistame $d_0 := D_{\mathrm{KL}}[q_u^{(0)} \times q_x^{(0)} || p]$ ja $d_\infty := \min_{q_u \in S_u, q_x \in S_x} L(q_u, q_x)$. Illustratiivselt võime ette kujutada olukorda, kus iteratsiooni sammul i paraneb KL kaugus $\frac{1}{2^{i+1}}(d_0 - d_\infty)$ võrra. Selline olukord ei oleks lemma 1 väitega vastuolus. Kuna kõikide KL kauguste paranemiste summa on $\sum_{i=1}^{\infty} \frac{1}{2^{i+1}}(d_0 - d_\infty) = \frac{d_0 - d_\infty}{2}$, siis kehtib koondumine $d_k \xrightarrow{k} \frac{d_0 + d_\infty}{2}$, mis ei ole koonduvus püsipunktini, sest vastaval juhul kehtiks $d_k \xrightarrow{k} d_\infty$. Täiendavalt, kui ei ole koondumist $d_k \xrightarrow{k} d_\infty$, siis me ei saa lihtsalt väita, et jada $\left(q_u^{(k)}, q_x^{(k)}\right)$ koondub.

Enne teoreemini jõudmist, tõestame veel ühe abistava lemma, mis väidab, et iteratsiooni sammu tulemusena liiguvad simpleksi rajal paiknevad punktid simpleksi sisemusse.

Lemma 2. *Kui $S_u \times S_x \subseteq \operatorname{supp} p$, siis fikseeritud q_u korral ei ole*

$$\tilde{q}_x := \arg\min_{q_x} D_{\mathrm{KL}}\left[q_u \times q_x \| p\right]$$

simpleksi S_x rajapunkt ehk kehtib $\tilde{q}_x \in S_x^\circ$.

Tõestus. Oletame vastuväiteliselt, et $\tilde{q}_x \in \partial S_x$. Siis leidub selline $\boldsymbol{a} \in \mathcal{X}^T$, et $\tilde{q}_x(\boldsymbol{a}) = 0$. Vaatleme sellist jaotust $q_\lambda \in S_x$, kus $q_\lambda(\boldsymbol{a}) = \lambda$ ning kui $\boldsymbol{x} \neq \boldsymbol{a}$, siis $q_\lambda(\boldsymbol{x}) = (1-\lambda)\tilde{q}_x(\boldsymbol{x})$. Leiame λ, mis minimiseeriks KL kaugust

$$D_{\mathrm{KL}}\left[q_u \times q_\lambda \| p\right].$$

Leiame KL kauguse tuletise muutuja λ järgi ja saame

$$\begin{aligned} 0 &= \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln \frac{\lambda q_u(\boldsymbol{u})}{p(\boldsymbol{a}, \boldsymbol{u})} - \sum_{\boldsymbol{x} \neq \boldsymbol{a}, \boldsymbol{u}} \tilde{q}_x(\boldsymbol{x}) q_u(\boldsymbol{u}) \ln \frac{(1-\lambda)\tilde{q}_x(\boldsymbol{x}) q_u(\boldsymbol{u})}{p(\boldsymbol{x}, \boldsymbol{u})} \\ &= \ln \frac{\lambda}{1-\lambda} + \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln \frac{q_u(\boldsymbol{u})}{p(\boldsymbol{a}, \boldsymbol{u})} - \sum_{\boldsymbol{x} \neq \boldsymbol{a}, \boldsymbol{u}} q_u(\boldsymbol{u}) \tilde{q}_x(\boldsymbol{x}) \ln \frac{\tilde{q}_x(\boldsymbol{x}) q_u(\boldsymbol{u})}{p(\boldsymbol{x}, \boldsymbol{u})} \end{aligned}$$

ehk $\ln \frac{\lambda}{1-\lambda}$ on tõkestatud, sest iga $\boldsymbol{u} \in S_u, \boldsymbol{x} \in S_x$ korral $p(\boldsymbol{u}, \boldsymbol{x}) > 0$. Seepärast ka optimaalne $\lambda \in (0, 1)$ ehk $q_\lambda \in S_x^\circ$ ning

$$D_{\mathrm{KL}}\left[q_u \times q_\lambda \| p\right] < D_{\mathrm{KL}}\left[q_u \times \tilde{q}_x \| p\right]$$

ning oleme saanud vastuolu eeldusega, et $\tilde{q}_x$ on minimaalne. □

Teoreem 1. *Iga initsialisatsiooni $q_u^{(0)}$ ja $q_x^{(0)}$ korral ruumis $S_u \times S_x$ annab algoritmi (1.7), (1.8) piirelement ülesande (1.6) lahendi. Täpsemalt, toimub koondumine*

$$\left(q_u^{(k)}, q_x^{(k)}\right)_k \xrightarrow{k} \left(q_u^{(\infty)}, q_x^{(\infty)}\right)$$

nii, et

$$L\left(q_u^{(\infty)}, q_x^{(\infty)}\right) = \min_{q_u \in S_u, q_x \in S_x} L(q_u, q_x).$$

Tõestus. Oletame vastuväiteliselt, et leidub reaalarv $a > d_\infty$ ja initsialisatsioon $q_u^{(0)}, q_x^{(0)}$ nii, et

$$L\left(q_u^{(k)}, q_x^{(k)}\right) \xrightarrow{k} a > d_\infty. \tag{1.9}$$

Et ruum $S_u \times S_x$ on ruumi $\mathbb{R}^{|\mathcal{U}|^T|\mathcal{X}|^T}$ kompaktne alamruum, siis iga jada selles ruumis sisaldab koonduva alamjada. Seega saame vaadelda sellist osajada

$$\left(q_u^{(k_j)}, q_x^{(k_j)}\right)_j,$$

mis koondub, ütleme elemendiks $(\tilde{q}_u, \tilde{q}_x)$, kusjuures $L(\tilde{q}_u, \tilde{q}_x) = a$, sest koonduva jada (d_k) iga osajada koondub samaks elemendiks.

Et $(\tilde{q}_u, \tilde{q}_x)$ ei ole eelduse järgi miinimumkoht, siis lemma 1 põhjal saame algoritmi sammu järel paari, millel on väiksem KL kaugus. Üldisust kitsendamata olgu selleks paar $(\tilde{q}'_u, \tilde{q}_x)$.

Tähistame $\Delta := \tilde{q}'_u - \tilde{q}_u$ ja $\varepsilon := L(\tilde{q}_u, \tilde{q}_x) - L(\tilde{q}'_u, \tilde{q}_x)$. Et järgnevas võib juhtuda, et $q_u^{(k_j)} + \Delta \notin S_u$, ei ole see probleem, sest lemma 2 põhjal $\tilde{q}'_u$ on simpleksi S_u sisepunkt ja seega koonduvusest $q_u^{(k_j)} + \Delta \xrightarrow{j} \tilde{q}'_u$ järeldub, et leidub $J \in \mathbb{N}$ nii, et iga $j > J$ korral kehtib $q_u^{(k_j)} + \Delta \in S_u$.

Funktsionaali L pidevusest ja koondumisest $\left(q_u^{(k_j)} + \Delta, q_x^{(k_j)}\right) \underset{j}{\to} (\tilde{q}'_u, \tilde{q}_x)$ järeldub, et iga $0 < \varepsilon_0 < \varepsilon$ korral leidub selline $N(\varepsilon_0)$ nii, et iga $j \geq N(\varepsilon_0)$ korral

$$\left|L\left(q_u^{(k_j)} + \Delta, q_x^{(k_j)}\right) - L\left(\tilde{q}'_u, \tilde{q}_x\right)\right| < \varepsilon_0.$$

Sestap iga $j \geq \max\{N(\varepsilon_0), J\}$ korral

$$\begin{aligned} L(\tilde{q}_u, \tilde{q}_x) - L\left(q_u^{(k_j)} + \Delta, q_x^{(k_j)}\right) &= \left[L(\tilde{q}_u, \tilde{q}_x) - L(\tilde{q}'_u, \tilde{q}_x)\right] + \left[L(\tilde{q}'_u, \tilde{q}_x) - L\left(q_u^{(k_j)} + \Delta, q_x^{(k_j)}\right)\right] \\ &\geq \varepsilon - \varepsilon_0 > 0 \end{aligned}$$

ning $q_u^{(k_j)} + \Delta \in S_u$ ehk $q_u^{(k_j)} + \Delta, q_x^{k_j}$ defineerivad korrektse tõenäosusmõõdu ja seega

$$L\left(q_u^{(k_j)} + \Delta, q_x^{(k_j)}\right) = L(\tilde{q}_u, \tilde{q}_x) - \varepsilon + \varepsilon_0 < a.$$

Sellega oleme näidanud, et kui $j \geq \max\{N(\varepsilon_0), J\}$, siis

$$\begin{aligned} L\left(q_u^{(k_{j+1})}, q_x^{(k_j)}\right) &\leq L\left(q_u^{(k_j)} + \Delta, q_x^{(k_j)}\right) \\ &< a, \end{aligned}$$

kus esimene võrratus tuleb (1.7) minimaalsusest. Jada

$$\left(L\left(q_u^{(k)}, q_x^{(k)}\right)\right)_k$$

monotoonse mittekasvavuse põhjal saame ka, et kui $k > k_{\max\{N(\varepsilon_0),J\}}$, siis

$$L\left(q_u^{(k)}, q_x^{(k)}\right) \leq L\left(q_u^{(k_{N(\varepsilon_0)})}, q_x^{(k_{N(\varepsilon_0)})}\right) < a.$$

Oleme jõudnud vastuoluni eeldusega (1.9). Sellega oleme näidanud, et iga initsialisatsiooni $q_u^{(0)}, q_x^{(0)}$ korral koondub algoritm globaalsesse miinimumi

$$L\left(q_u^{(k)}, q_x^{(k)}\right) \xrightarrow{k} d_\infty.$$

Nüüd näitame, et jadal $\left(q_u^{(k)}, q_x^{(k)}\right)_k$ leidub piirväärtus. Rangelt kumerusest järeldub, et $\left(q_u^{(\infty)}, q_x^{(\infty)}\right) := L^{-1}(d_\infty)$ on unikaalne. Igal koonduva alamjada $\left(q_u^{(k_j)}, q_x^{(k_j)}\right)_j$ korral kehtib $L\left(q_u^{(k_j)}, q_x^{(k_j)}\right) \xrightarrow{j} d_\infty$ ja seega L pidevuse ja $L^{-1}(d_\infty)$ unikaalsuse tõttu $\left(q_u^{(k_j)}, q_x^{(k_j)}\right) \xrightarrow{j} \left(q_u^{(\infty)}, q_x^{(\infty)}\right)$. Kuna hulga $S_u \times S_x$ kompaktsuse tõttu on igal alamjadal koonduv (alam)alamjada, mis eelneva põhjal koondub samaks punktiks, siis originaalne jada koondub ka samaks punktiks $\left(q_u^{(k)}, q_x^{(k)}\right) \xrightarrow{k} \left(q_u^{(\infty)}, q_x^{(\infty)}\right)$. □

Kokkuvõttes oleme näidanud, et ülesandel (1.3) on täpselt üks lahend ning sammudest

(1.7) ja (1.8) koosnev iteratsioonimeetod koondub selleks lahendiks.

Nüüd uurime kuidas arvutada iteratsioonimeetodi samme (1.7) ja (1.8). Üldisust kaotamata arvutame sammu (1.7) ehk leiame

$$\arg\min_{q_u \in \mathcal{P}(\mathcal{U})} D_{\mathrm{KL}}[q_u \times q_x^{(i)} || P].$$

Selleks fikseeritud i korral tähistame $q_x := q_x^{(i)}$ ning minimiseerime funktsionaali $q_u \mapsto D_{\mathrm{KL}}[q_u \times q_x || p]$:

$$\begin{aligned} D_{\mathrm{KL}}[q_u \times q_x || p] &= \sum_{\boldsymbol{u},\boldsymbol{x}} q_u(\boldsymbol{u}) q_x(\boldsymbol{x}) \ln\left(\frac{q_u(\boldsymbol{u}) q_x(\boldsymbol{x})}{p(\boldsymbol{u},\boldsymbol{x})}\right) \\ &= -H[q_x] + \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln q_u(\boldsymbol{u}) - \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \sum_{\boldsymbol{x}} q_x(\boldsymbol{x}) \ln p(\boldsymbol{u},\boldsymbol{x}). \end{aligned}$$

Nüüd defineerime iga $\boldsymbol{u}$ korral

$$q^*(\boldsymbol{u}) := \exp\left(\sum_{\boldsymbol{x}} q_x(\boldsymbol{x}) \ln p(\boldsymbol{u},\boldsymbol{x})\right),$$

mida kasutades saame kirjutada

$$\begin{aligned} D_{\mathrm{KL}}[q_u \times q_x || p] &= -H[q_x] + \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln q_u(\boldsymbol{u}) - \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln q^*(\boldsymbol{u}) \\ &= -H[q_x] + \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln \frac{q_u(\boldsymbol{u})}{q^*(\boldsymbol{u})}. \end{aligned}$$

Paneme tähele, et iga $\boldsymbol{u}$ korral $q^*(\boldsymbol{u}) \in [0,1]$. Üldiselt ei ole mõõt q^* tõenäosusmõõt. Küll aga me varasemalt eeldasime, et simpleksite otsekorrutis $S_u \times S_x$, mis on alamhulk ühisjaotuse p kandjal, ei ole tühihulk, seega iga $\boldsymbol{x}$ korral, kus $q_x(\boldsymbol{x}) > 0$, kehtib ka iga $\boldsymbol{u}$ korral $p(\boldsymbol{u},\boldsymbol{x}) > 0$. Et q_x on tõenäosusmõõt, siis selline $\boldsymbol{x}$ leidub ja leidub ka $\boldsymbol{u}$ nii, et $q^*(\boldsymbol{u}) > 0$. Seega on võimalik mõõtu q^* normaliseerida.

Defineerime $Z := \sum_{\boldsymbol{u}} q_u^*(\boldsymbol{u}) > 0$ ning $\bar{q}_u(\boldsymbol{u}) = 1/Z\ q_u^*(\boldsymbol{u})$, kus $\bar{q}_u$ on tõenäosusmõõt. Kirjutame KL kauguse kui

$$D_{\mathrm{KL}}[q_u \times q_x || p] = -H[q_x] + \sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln \frac{q_u(\boldsymbol{u})}{Z\bar{q}(\boldsymbol{u})} = -H[q_x] - \ln Z + D_{\mathrm{KL}}[q_u || \bar{q}_u],$$

kusjuures jaotus q_x, konstant Z ja $\bar{q}_u$ ei sõltu mõõdust q_u, seega minimiseerimine on samaväärne $D_{\mathrm{KL}}[q_u||\bar{q}_u]$ minimiseerimisega. Et KL kaugus on mittenegatiivne ja 0 parajasti siis, kui mõõdud on võrdsed, siis minimiseeriv jaotus ongi $\bar{q}_u$. Saame kirjutada tulemuse $\ln q_u^{(i+1)} = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln p(\boldsymbol{u}, \boldsymbol{x}) - \ln Z_u^{(i+1)}$ või

$$q_u^{(i+1)}(\boldsymbol{u}) = \frac{1}{Z_u^{(i+1)}} \exp\left(\sum_{\boldsymbol{x}} q_x(\boldsymbol{x}) \ln p(\boldsymbol{u}, \boldsymbol{x}) \right),\ Z_u^{(i+1)} = \sum_{\boldsymbol{u}} \exp\left(\sum_{\boldsymbol{x}} q_x(\boldsymbol{x}) \ln p(\boldsymbol{u}, \boldsymbol{x}) \right). \tag{1.10}$$

Nüüd näitame, et saame sarnaselt arutleda ka mõõtude $q_1, \ldots, q_T$ korral, kus iga q_t on tõenäosusmõõt hulgal $\mathcal{U} \times \mathcal{X}$. Defineerime $q := q_1 \times \ldots \times q_T$ ning minimiseerime KL kaugust $D_{\mathrm{KL}}[q\|p]$. Seda teeme parandades mõõte igal iteratsioonil ükshaaval $t = 1, \ldots, T$ korral. Saame kirjutada mõõdu q_t paranduse analoogselt valemile (1.10). Vaatleme mõõtude $q_u \times q_x$ asemel korrutismõõtu $q_t \times q_{\setminus t}$, kus

$$q_{\setminus t} := q_1 \times \ldots \times q_{t-1} \times q_{t+1} \times \ldots q_T,$$

ning olgu simpleksid $S_1, \ldots, S_T$ sellised, et $S_1 \times \ldots \times S_T \subseteq \operatorname{supp} p$. Vaatleme iteratsioone, kus iga $t = 1, \ldots, T$ korral

$$q_t^{(i+1)}(u_t, x_t) = \frac{1}{Z_t^{(i+1)}} \exp\left(\sum_{\boldsymbol{u}_{\setminus t}, \boldsymbol{x}_{\setminus t}} q_{\setminus t}(\boldsymbol{u}_{\setminus t}, \boldsymbol{x}_{\setminus t}) \ln p(\boldsymbol{u}, \boldsymbol{x}) \right), \tag{1.11}$$

$$Z_t^{(i+1)} = \sum_{u_t, x_t} \exp\left(\sum_{\boldsymbol{u}_{\setminus t}, \boldsymbol{x}_{\setminus t}} q_x(\boldsymbol{x}) \ln p(\boldsymbol{u}, \boldsymbol{x}) \right).$$

Saame algoritmi, mille iteratsioonidele vastav KL kauguste jada koondub miinimumiks

$$d_k = L\left(\prod_{t=1}^{T} q_t^{(k)}\right) \underset{k}{\to} \min_{q \in \prod_{t=1}^{T} S_t} L(q),$$

kus

$$L : S_1 \times \ldots \times S_T \to \mathbb{R}, \ \left(\prod_{t=1}^{T} q_t\right) \mapsto D_{\mathrm{KL}}\left[\prod_{t=1}^{T} q_t \| p\right].$$

Et L on rangelt kumer, on miinimumkoht üheselt määratud ning argumendid koonduvad lahendiks $(q_1^k, \ldots, q_T^k) \underset{k}{\to} (q_1^\infty, \ldots, q_T^\infty)$.

1.4 Marginaalide lähendamine

Uurime, kas meil on võimalik lahendada minimiseerimisülesannet (1.4), kus argumendid on ära vahetatud ja me lahendame marginaali p_x lähendi q_x leidmiseks ülesannet

$$\arg\min_{q_u \times q_x} D_{\mathrm{KL}}[p \| q_u \times q_x]. \tag{1.12}$$

Saab näidata, et lahendite leidmine on raske ning lihtsamad iteratiivsed algoritmid ei koondu lahenditeks. Näitame, et lahendiks $q_u \times q_x$ ongi p marginaaljaotuste korrutis $p_u \times p_x$. Selleks kirjutame

$$\begin{aligned} D_{\mathrm{KL}}[p||q_u \times q_x] &= \sum_{\boldsymbol{u},\boldsymbol{x}} p(\boldsymbol{u},\boldsymbol{x}) \left(\ln p(\boldsymbol{u},\boldsymbol{x}) - \ln q_u(\boldsymbol{u}) - \ln q_x(\boldsymbol{x})\right) \\ &= \sum_{\boldsymbol{u},\boldsymbol{x}} p(\boldsymbol{u},\boldsymbol{x}) \ln p(\boldsymbol{u},\boldsymbol{x}) - \sum_{\boldsymbol{u}} p_u(\boldsymbol{u}) \ln q_u(\boldsymbol{u}) - \sum_{\boldsymbol{x}} p_x(\boldsymbol{x}) \ln q_x(\boldsymbol{x}) \end{aligned}$$

ja otsides minimiseerivaid q_x, q_u näeme, et saame seda teha eraldi. Kirjutades ainukese mõõdust q_x sõltuva liikme välja ja kasutades Gibbsi võrratust (Lember, J., 2022) saame, et

$$-\sum_{\boldsymbol{u},\boldsymbol{x}} p(\boldsymbol{u},\boldsymbol{x}) \ln q_x(\boldsymbol{x}) = -\sum_{\boldsymbol{x}} p_x(\boldsymbol{x}) \ln q_x(\boldsymbol{x}) \geq -\sum_{\boldsymbol{x}} p_x(\boldsymbol{x}) \ln p_x(\boldsymbol{x}).$$

Võrratuses kehtib võrdus parajasti siis, kui mõõdud q_x ja p_x on võrdsed. See tähendab,

et ülesande (1.12) lahendisse kuuluv mõõt q_x on marginaalmõõt p_x. Analoogselt saame, et lahendisse kuuluv mõõt q_u on marginaalmõõt p_u.

Kuigi marginaalmõõtude leidmine on teoreetiliselt võimalik, ei ole need üldiselt meie mudelites Markovi omadusega, mispärast ei saa Viterbi raja leidmiseks rakendada Viterbi algoritmi. Nagu me peatükis 1.2 mainisime, on marginaalmõõdust p_x Viterbi raja leidmine NP-raske.

Kuigi ülesande (1.12) lahendamine on õigem, siis seda on teha raske. Ülesande (1.4) lahend q_x ei ole marginaal p_x, kuid lahendi q_x Viterbi rada võiks mingitel juhtudel olla hea kandidaat Viterbi rajale (1.3), sest KL kauguse (1.5) minimiseerimine tähendab kahe jaotuse $q_u \times q_x$ ja p lähendamist. Seda soovime töö eksperimentaalosas lähemalt uurida.

2 Algoritmid

Kirjeldame algoritme, mida kasutame hilisemates peatükkides. Vaatleme tõenäosusmõõtu q_x, kus

$$q_x(\boldsymbol{x}) := \pi_x(x_1) \prod_{t=2}^{T} q_t(x_t|x_{t-1}), \tag{2.1}$$

kusjuures mõõdud π_x, $q_t(\cdot|x_{t-1})$ ei pruugi olla tõenäosusmõõdud; viimane võib isegi olla mingi t, x_{t-1} korral nullmõõt. Tähistame fikseeritud $i < j$ korral vektoreid kui $\boldsymbol{x}_{i:j} = (x_i, \ldots, x_j)$, $\boldsymbol{x} = (x_t)_{t=1}^T$ ning $\boldsymbol{x}_{\setminus i,j} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_T)$. Edasi-tagasi algoritm kirjeldab, kuidas efektiivselt leida marginaali $q_x(x_t) := \sum_{x_{\setminus t}} q_x(\boldsymbol{x}_{1:T})$ ja ühismarginaali $q_x(x_{t-1}, x_t) := \sum_{x_{\setminus t-1,t}} q_x(\boldsymbol{x}_{1:T})$. Viterbi algoritmi abil leiame mõõdu q_x jaoks Viterbi raja.

2.1 Edasi-tagasi algoritm

Otsime marginaali $q_x(x_t)$, selleks kirjutame

$$\begin{aligned} q_x(x_t) &= \sum_{x_{1:t-1}} \sum_{x_{t+1:T}} \pi_x(x_1) \prod_{t=2}^{T} q_t(x_t|x_{t-1}) \\ &= \left(\sum_{x_{1:t-1}} \pi_x(x_1) \prod_{t'=2}^{t-1} q_{t'}(x_{t'}|x_{t'-1}) \right) \cdot \left(\sum_{x_{t+1:T}} \prod_{t'=t+1}^{T} q_{t'}(x_{t'}|x_{t'-1}) \right) \\ &= \alpha_t(x_t) \cdot \beta_t(x_t), \end{aligned} \tag{2.2}$$

kus kirjutame $\alpha_1(x_1) = \pi(x_1)$, $\beta_T(x_T) = 1$ ja muidu $t \in \{1, \ldots, T\}$ korral

$$\begin{aligned} \alpha_t(x_t) :&= \sum_{x_{1:t-1}} p(\boldsymbol{x}_{1:t}) = \sum_{x_{1:t-1}} \pi_x(x_1) \prod_{t'=2}^{t-1} q_{t'}(x_{t'}|x_{t'-1}), \\ \beta_t(x_t) :&= \sum_{x_{t+1:T}} p(\boldsymbol{x}_{t+1:T}|x_t) = \sum_{x_{t+1:T}} \prod_{t'=t+1}^{T} q_{t'}(x_{t'}|x_{t'-1}). \end{aligned}$$

Ilmselt saame kirjutada rekursiivselt iga $t > 1$ ja α_t jaoks ning iga $t < T$ ja β_t jaoks

$$\alpha_t(x_t) = \sum_{x_{t-1}} \alpha_t(x_{t-1}) q_t(x_t|x_{t-1}) \text{ ja}$$
$$\beta_t(x_t) = \sum_{x_{t+1}} q_{t+1}(x_{t+1}|x_t)\beta_{t+1}(z_{t+1}).$$

Analoogselt saame kirjutada ühismarginaali $q_x(x_{t-1}, x_t)$ kui

$$\begin{aligned} q_x(x_{t-1}, x_t) &= \sum_{x_{1:t-2}} \sum_{x_{t+1:T}} \pi_x(x_1) \prod_{t=2}^{T} q_t(x_t|x_{t-1}) \\ &= \left(\sum_{x_{1:t-2}} \pi_x(x_1) \prod_{t'=2}^{t-1} q_{t'}(x_{t'}|x_{t'-1}) \right) q_t(x_t|x_{t-1}) \left(\sum_{x_{t+1:T}} \prod_{t'=t+1}^{T} q_{t'}(x_{t'}|x_{t'-1}) \right) \\ &= \alpha_{t-1}(x_t) q_t(x_t|x_{t-1}) \beta_t(x_t). \end{aligned} \tag{2.3}$$

2.2 Viterbi algoritm

Vaatleme tõenäosusmõõtu q_x kujul (2.1), kus

$$q_x(\boldsymbol{x}) := \pi_x(x_1) \prod_{t=2}^{T} q_t(x_t|x_{t-1}),$$

sooviga leida Viterbi rada $\boldsymbol{x}^* \in \mathcal{X}^T$, kus $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} q_x(\boldsymbol{x})$. Paneme tähele, et me saame leida selle ajalise keerukusega $\mathcal{O}(|\mathcal{X}|^2 T)$ järgmiselt. Defineerime funktsiooni δ, kus $\delta(x_1) = q(x_1)$ ja $t > 1$ korral

$$\delta(x_t) = \max_{x_{1:t-1}} q_x(\boldsymbol{x}_{1:t}),$$

mille saame samaväärselt kirjutada iga $t = 2, \ldots, T$ korral

$$\delta(x_t) = \max_{x_{t-1}} q(x_t|x_{t-1})\delta(z_{t-1}),$$

kus jätame meelde ka

$$c(z_t) = \arg\max_{z_{t-1}} p(z_t|z_{t-1})b(x_{t-1}),$$

kusjuures implementeerides võib iga t korral ära unustada kõik $\delta(x_{t'})$, kus $t' < t-1$.

Seejärel võtame $x_T^* = \arg\max_{x_T} \delta(x_T)$ ning leiame Viterbi raja leides iga $t = T-1, \ldots, 1$ korral $x_t^* = c(x_{t+1})$.

3 Viterbi raja lähendamine

Nüüdseks oleme tutvunud paarikaupa Markovi ahelatel (1.2) Viterbi raja probleemiga (1.3) ja õigustanud iteratiivset lahendusviisi peatükkides 1.3, 1.4, mille abil luua algoritm ülesande lähendi leidmiseks. Eelmises peatükis tutvustasime tuntud algoritme meie ülesande kontekstis ning nüüd implementeerime kaks algoritmi.

Mõlema puhul vaatleme diskreetset mittehomogeenset paarikaupa Markovi ahelat $\{U_t, X_t\}$ jaotusega p, mida kirjeldab 1.2, kus algjaotus on π ning üleminekumaatriks ajahetkel t on p_t ehk paaride $(u_{t-1}, x_{t-1}), (u_t, x_t) \in \mathcal{U}\times\mathcal{X}$ korral on üleminekutõenäosus $p_t(u_t, x_t|u_{t-1}, x_{t-1})$. Soov on leida jaotus q, mille abil leida Viterbi raja lähend $\boldsymbol{x}^*$. Esimeses algoritmis eeldame sõltumatust $q(\boldsymbol{u}, \boldsymbol{x}) = q_u(\boldsymbol{u})q_x(\boldsymbol{x})$ ning teises $q(\boldsymbol{u}, \boldsymbol{x}) = \prod_{t=1}^{T} q_t(u_t, x_t)$.

3.1 Belief propagation (BP)

Me oleme näidanud, milline on tõenäosusmõõdu

$$q_u \times q_x : \mathcal{U}^T \times \mathcal{X}^T \ni (\boldsymbol{u}, \boldsymbol{x}) \mapsto q_u(\boldsymbol{u})q_x(\boldsymbol{x}) \in [0, \infty)$$

puhul iteratiivse algoritmi samm (1.10) minimiseerimaks KL kaugust $D_{\mathrm{KL}}[q_u \times q_x \| p]$. Näitame, kuidas käib sellise algoritmi implementeerimine, kui p on mittehomogeenne diskreetne paarikaupa Markovi ahel.

Vaatleme tõenäosusmõõtu $q_u^{(0)}$, kus iga vektori $\boldsymbol{u} \in \mathcal{U}^T$ korral kehtib (2.1) ehk

$$q(\boldsymbol{u})^{(0)} = \pi_u^{(0)}(u_1) \prod_{t=2}^{T} \tilde{q}_t^{(0)}(u_t|u_{t-1}),$$

kusjuures mõõt $\tilde{q}_t^{(0)}(\cdot|u_{t-1})$ ei pruugi olla tõenäosusmõõt. Aga meenutame, et me oskame leida $q_u^{(0)}(u_t)$ iga t korral ja $q_u^{(0)}(u_{t-1}, u_t)$ iga $t > 2$ korral (2.2) ja (2.3) abil. Kirjutame

Algorithm 1: BP algoritmi pseudokood

$i = 0$
$q_u^{(i)}, q_x^{(i)} = \text{init}()$
while $D_{\text{KL}}\left[q_u^{(i)} \times q_x^{(i)} \| p\right]$ *not converged* **do**
  for $t = 1, \ldots, T$ **do**
    if $t = 1$ **then**
      for each u_t **do** $q_u^{(i)}(u_t) = \text{forwBack}(q_u^{(i)}, t)$
    else
      for each u_t, u_{t-1} **do** $q_u^{(i)}(u_t), q_u^{(i)}(u_t, u_{t-1}) = \text{forwBack}(q_u^{(i)}, t)$
  $\ln \pi_x^{(i)}(x_1) = \sum_{u_1 \in \mathcal{U}} q_u^{(i)}(u_1) \ln p(u_1, x_1)$
  for $t = 2, \ldots, T$ **do**
    for each x_t, x_{t-1} **do**
      $\ln q_x^{(i)}(x_t | x_{t-1}) = \sum_{u_{t-1,t} \in \mathcal{U}^2} q_u^{(i)}(u_{t-1}, u_t) \ln p_t(u_t, x_t | u_{t-1}, x_{t-1})$
  for each $\boldsymbol{x}$ **do** $\tilde{q}_x^{(i)}(\boldsymbol{x}) = \pi_x^{(i)}(x_1) \prod_{t=2}^{T} q_x^{(i)}(x_t | x_{t-1})$
  $Z_x^{(i)} = \sum_{\boldsymbol{x}} \tilde{q}_x^{(i)}(\boldsymbol{x})$
  for each $\boldsymbol{x}$ **do** $q_x^{(i)}(\boldsymbol{x}) = \tilde{q}_x^{(i)}(\boldsymbol{x}) / Z_x^{(i)}$
  $q_u^{(i+1)} =$ **repeat** *4-14* **swapping** $q_u^{(i)}$ **for** $q_x^{(i)}$
  $i{+}{+}$
return *Viterbi*$(q_x^{(i)})$

variatsioonilise Bayesi iteratsiooni vastavalt tulemusele (1.10) ja saame

$$\ln q_x^{(0)}(\boldsymbol{x}) = \sum_{\boldsymbol{u}} \ln p(\boldsymbol{u}, \boldsymbol{x}) q_u^{(0)}(\boldsymbol{u}) + \ln Z_x^{(0)} \tag{3.1}$$

ning näitame, et tõenäosusmõõt $q_x^{(0)}$ on võimalik samuti esitada sarnaselt mõõdule q_u kujul (2.1). Uurime eelmise avaldise normaliseerimata paremat poolt

$$\begin{aligned}\sum_{\boldsymbol{u}} \ln p(\boldsymbol{u}, \boldsymbol{x}) q_u^{(0)}(\boldsymbol{u}) &= \sum_{\boldsymbol{u}} \ln \left(\pi(x_1, u_1) \prod_{t=2}^{T} p_t(x_t, u_t | x_{t-1}, u_{t-1}) \right) q_u^{(0)}(\boldsymbol{u}) \\ &= \sum_{u_1} \ln \pi(x_1, u_1) q_u^{(0)}(u_1) + \sum_{t=2}^{T} \sum_{\boldsymbol{u}_{t-1,t}} \ln p_t(u_t, x_t | u_{t-1}, x_{t-1}) q_u^{(0)}(u_{t-1}, u_t).\end{aligned}$$

Defineerime iga t ning iga $x_t \in \mathcal{X}$ korral $\ln \pi_x^{(0)}(x_1)$ ja $\ln q_x^{(0)}(x_t|x_{t-1})$, kus

$$\ln \pi_x^{(0)}(x_1) := \sum_{u_1 \in \mathcal{U}} q_u^{(0)}(u_1) \ln p(u_1, x_1)$$

ning iga $t > 1$ ja iga $(x_{t-1}, x_t) \in \mathcal{X}^2$ jaoks

$$\ln q_x^{(0)}(x_t|x_{t-1}) := \sum_{u_{t-1,t} \in \mathcal{U}^2} q_u^{(0)}(u_{t-1}, u_t) \ln p_t(u_t, x_t|u_{t-1}, x_{t-1}).$$

Oleme saanud, et

$$q_x^{(0)}(\boldsymbol{x}) \propto \pi_x^{(0)}(x_1) \prod_{t=2}^{T} q_x^{(0)}(x_t|x_{t-1}).$$

Avaldise parem pool ei ole üldiselt tõenäosusmõõt, aga eelduse kohaselt $\operatorname{supp} q_u \times q_x \subseteq \operatorname{supp} p$, seega peatüki 1.3 põhjal ei ole see nullmõõt. Defineerime

$$\tilde{q}_x^{(0)}(\boldsymbol{x}) = \pi_x^{(0)}(x_1) \prod_{t=2}^{T} q_x^{(0)}(x_t|x_{t-1}).$$

Järgmisena keskendume mõõdu $\tilde{q}_x^{(0)}$ normaliseerimisele ehk leiame avaldises (3.1) suuruse $Z_x^{(0)}$. Praktiline on leida suurus $Z_x^{(0)}$ mõõdu $\tilde{q}_x^{(0)}$ edasi-tagasi algoritmi $\tilde{\alpha}, \tilde{\beta}$ komponentide abil, sest nii normaliseerimine kui ka $q_x^{(0)}(x_t), q_x^{(0)}(x_{t-1}, x_t)$ marginaalide leidmine on kergesti tehtavad nende komponentide abil. On lihtne näha, et saame kirjutada tõenäosusmõõdu $q_x^{(0)}$ edasi-tagasi komponendid kui

$$\alpha_t(x_t) = \tilde{\alpha}_t(x_t)/Z_x^{(0)}.$$

Leiame mõõdu $\tilde{q}_x^{(0)}$ jaoks iga $t \in \{1, \ldots, T\}$ jaoks $\tilde{\alpha}_t, \tilde{\beta}_t$ ja saame näiteks $t = 1$ korral

$$Z_x^{(0)} = \sum_{x_t} \tilde{\alpha}_1(x_1)\tilde{\beta}_1(x_1). \tag{3.2}$$

Saame tõenäosusmõõdu

$$q_x^{(0)}(\boldsymbol{x}) = \frac{1}{Z_x^{(0)}} \pi_x^{(0)}(x_1) \prod_{t=2}^{T} q_x^{(0)}(x_t|x_{t-1}). \tag{3.3}$$

Nüüd iteratsiooni teises pooles leiame

$$q^{(1)}(\boldsymbol{u}) = \frac{1}{Z_u^{(1)}} \pi_u^{(1)}(u_1) \prod_{t=2}^{n} q_u^{(1)}(u_t|u_{t-1}),$$

kus

$$\ln \pi_u^{(1)}(u_1) := \sum_{x_1 \in \mathcal{X}} \pi_x^{(0)}(x_1) \ln \pi(x_1, u_1)$$
$$\ln q_u^{(1)}(u_t|u_{t-1}) := \sum_{\boldsymbol{x}_{t-1,t} \in \mathcal{X}^2} q_x^{(0)}(x_{t-1}, x_t) \ln p_t(x_t, u_t|x_{t-1}, u_{t-1})$$

ja $Z_u^{(1)}$ on mõõdu

$$\pi_u^{(1)}(u_1) \prod_{t=2}^{n} q_u^{(1)}(u_t|u_{t-1})$$

edasi tagasi komponentide abil leitav normaliseeriv konstant analoogselt avaldisele (3.2).

Iteratsiooni teist poolt saime teha analoogselt, sest kui $q_u^{(0)}$ on kujul (2.1), siis on ka (3.3) põhjal $q_x^{(0)}$ ja $q_u^{(1)}$ Markovi ahelad jne. Seega sobivatel algväärtustustel, kus initisialiseerime $q_u^{(0)}$ Markovi ahelana, ei toimu optimiseerimine kitsendusega üle Markovi ahelate, vaid üle kõikide mõõtude $\mathcal{P}(\mathcal{X})$.

Peatükis 1.3 kirjeldatud ja siin peatükis implementeeritud iteratiivne algoritm tagab iteratsioonidel KL kaugust minimiseeriva korrutismõõdu $q_u \times q_x$.

Viterbi raja lähendi ehk lähendi ülesandele (1.3) anname kui Viterbi raja üle leitud mõõdu q_x

$$\arg\max_{\boldsymbol{x}} q_x(\boldsymbol{x}).$$

Loodud algoritm on hea, kui piisavalt paljude mittehomogeensete diskreetsete paarikaupa

Markovi ahelate $p(\boldsymbol{u}, \boldsymbol{x})$ korral

$$\arg\max_{\boldsymbol{x}} q_x(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} \sum_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{x}).$$

3.2 Variational Message Passing (VMP)

Algorithm 2: VMP algoritmi pseudokood

$i = 0$
$q_1^{(i)}, \ldots, q_T^{(i)} = \text{init}()$
while $D_{\text{KL}}\left[q_1^{(i)} \times \ldots \times q_T^{(i)} \| p\right]$ *not converged* **do**
  $i++$
  for $t = 1, \ldots, T$ **do**
    if *t=1* **then**
      for each u_t, x_t $\ln q_t^{*(i)}$**do**$(u_t, x_t) =$
       $\sum_{u_{t+1}, x_{t+1}} \ln p_{t+1}(u_{t+1}, x_{t+1} | u_t, x_t) q_{t+1}^{(i-1)}(u_{t+1}, x_{t+1})$
    else if *t=T* **then**
      for each u_t, x_t **do**
       $\ln q_t^*(u_t, x_t) = \sum_{u_{t-1}, x_{t-1}} \ln p_t(u_t, x_t | u_{t-1}, x_{t-1}) q_{t-1}^{(i)}(u_{t-1}, x_{t-1})$
    else
      for each u_t, x_t **do**
       $\ln q_t^*(u_t, x_t) = \sum_{u_{t+1}, x_{t+1}} \ln p_{t+1}(u_{t+1}, x_{t+1} | u_t, x_t) q_{t+1}^{(i-1)}(u_{t+1}, x_{t+1}) +$
       $\sum_{u_{t-1}, x_{t-1}} \ln p_t(u_t, x_t | u_{t-1}, x_{t-1}) q_{t-1}^{(i)}(u_{t-1}, x_{t-1})$
    $Z_t^{(i)} := \sum_{u_t, x_t} q_t^{*(i)}(u_t, x_t)$
    $q_t(u_t, x_t) = q_t^*(u_t, x_t) / Z_t^{(i)}$
for $t = 1, \ldots, T$ **do**
  $x_t^* = \arg\max_{x_t} \sum_{u_t} q_t(u_t, x_t)$
return $\boldsymbol{x}^*$

Vaatleme mittehomogeenset diskreetset paarikaupa Markovi ahelat $\{U_t, X_t\}_t$ jaotusega p (1.2).

Eeldame, et $q(\boldsymbol{u}, \boldsymbol{x}) = \prod_{t=1}^{T} q_t(u_t, x_t)$, kus iga t korral on $q_t(u_t, x_t)$ tõenäosusmõõt. Fik-

seerime nüüd t. Meid huvitav jaotus on $q_t(u_t, x_t)$ ning jaotuse

$$q_{\setminus t} = q_1 \times \ldots \times q_{t-1} \times q_{t+1} \times \ldots \times q_T$$

loeme fikseerituks.

Meenutame, et normaliseerimata iteratsioonisamm (1.11) on $t \in \{2, \ldots, T-1\}$ korral

$$\begin{aligned}
\ln \tilde{q}_t(u_t, x_t) &= \sum_{\boldsymbol{u}_{\setminus t}, \boldsymbol{x}_{\setminus t}} q_{\setminus t}(\boldsymbol{u}_{\setminus t}, \boldsymbol{x}_{\setminus t}) \ln p(\boldsymbol{u}, \boldsymbol{x}) \\
&= \sum_{u_1, x_1} q_1(u_1, x_1) \ln \pi(u_1, x_1) \\
&+ \sum_{\substack{u_2, x_2 \\ u_1, x_1}} q_1(u_1, x_1) q_2(u_2, x_2) \ln p_2(u_2, x_2 | u_1, x_1) \\
&+ \ldots \\
&+ \sum_{u_{t-1}, x_{t-1}} q_{t-1}(u_{t-1}, x_{t-1}) \ln p(u_t, x_t | u_{t-1}, x_{t-1}) \\
&+ \sum_{u_{t+1}, x_{t+1}} q_{t+1}(u_{t+1}, x_{t+1}) \ln p_{t+1}(u_{t+1}, x_{t+1} | u_t, x_t) \\
&+ \ldots \\
&+ \sum_{\substack{u_T, x_T \\ u_{T-1}, x_{T-1}}} q_{T-1}(u_{T-1}, x_{T-1}) q_T(u_T, x_T) \ln p_T(u_T, x_T | u_{T-1}, x_{T-1}).
\end{aligned}$$

Saame defineerida

$$\ln q_t^*(u_t, x_t) := \sum_{u_{t-1}, x_{t-1}} q_{t-1}(u_{t-1}, x_{t-1}) \ln p(u_t, x_t | u_{t-1}, x_{t-1}) \tag{3.4}$$

$$+ \sum_{u_{t+1}, x_{t+1}} q_{t+1}(u_{t+1}, x_{t+1}) \ln p_{t+1}(u_{t+1}, x_{t+1} | u_t, x_t), \tag{3.5}$$

kusjuures $\tilde{q}_t(u_t, x_t) \propto q_t^*(u_t, x_t)$ ning

$$\frac{q_t^*(u_t, x_t)}{\sum_{u_t, x_t} q_t^*(u_t, x_t)} = \frac{\tilde{q}_t(u_t, x_t)}{\sum_{u_t, x_t} \tilde{q}_t(u_t, x_t)}.$$

Kui $t = 1$ või $t = T$, siis vaatleme funktsiooni $\ln q_t^*$ defineerides vastavalt vaid teist (3.5) või esimest (3.4) liidetavat. Saame iteratsioonisammuks

$$\bar{q}_t(u_t, x_t) = \frac{q_t^*(u_t, x_t)}{Z_t}, \; Z_t := \sum_{u_t, x_t} q_t^*(u_t, x_t). \tag{3.6}$$

Algoritmi implementeerides saame initsialiseerida mõõdud $q_t^{(0)}$ ühtlase jaotusena üle tähestiku $\mathcal{U} \times \mathcal{X}$ iga t korral. Iteratsioonil i me uuendame (3.6) abil järjestikku $t = 1, \ldots, T$ korral mõõtu $q_t^{(i)}(u_t, x_t) = \bar{q}_t(u_t, x_t)$. Tasub tähele panna, et me kasutame $q_t^{(i)}$ leidmiseks mõõte $q_{t-1}^{(i)}$ ja $q_{t+1}^{(i-1)}$.

Kui algoritm on meie hinnangul koondunud, anname Viterbi raja lähendi pärast N iteratsiooni mõõtude $q_1^{(N)}, \ldots, q_T^{(N)}$ jaoks kui $(x_t^*)_{t=1}^T$, kus

$$x_t^* = \arg\max_{x_t} \sum_{u_t} q_t^{(N)}(u_t, x_t)$$

ning simulatsioonide abil saame võrrelda selle algoritmi käitumist peatükis 3.1 kirjeldatud algoritmiga.

4 Eksperimendid

Eksperimentide koostamiseks, mille abil uurida VMP ja BP algoritmide headust, on mitmeid viise. Me vaatleme eksperimentides paarikaupa varjatud tunnusega Markovi ahelaid.

Esimesel juhul uurime N mudelit ja andmestikku, mis on genereeritud võrdlemisi väikese $T \in \{10, 15, 20\}$ korral, kusjuures iga realisatsiooni on genereerinud erineva jaotusega mudel. Väikese T korral on võimalik täpselt leida, kui kaugele Viterbi rajast on jäänud töös kirjeldatud algoritmidega leitud rajad Hammingu kauguse mõttes ning mis on leitud rajade protsentiilid.

Kui Hammingu kaugus on enamasti väike, siis see annaks alust loota, et leides kõik rajad, mis on lähedal VMP ja BP abil leitud rajadele, on Viterbi rada nende hulgas. Veenvat teoreetilist põhjendust selle uskumiseks töö käigus ei otsitud, kuid suure T korral on vaja mingit strateegiat heade kandidaatradade otsimiseks, sest ei ole realistlik leida tõenäosused kõigile $|\mathcal{X}|^T$ rajale nt $T = 100$ korral. Toome siinkohal ka välja, et Hammingu ümbrust võib ka vaadata andmestiku genereerinud raja $x_1, \ldots, x_T$ korral, kuid on osutunud, et enamasti see rada on optimumist kaugel.

Teisel juhul vaatleme kaht erinevat eksperimenti vahelduva režiimiga mudeli puhul. Ka see on paarikaupa varjatud tunnusega Markovi ahel, kus vaatlusandmed on fikseeritud, kuid nõuame homogeenselt paarikaupa Markovi ahelalt rohkem struktuuri. Homogeense paarikaupa Markovi ahela puhul peame siinkohal silmas $\boldsymbol{U}, \boldsymbol{X}$ jaotust vaatlusandmeid teadmata. Vaatleme juhtu, kus iga realisatsiooni on genereerinud erinev mudel, ning viimasena juhtu, kus homogeenne paarikaupa Markovi ahela jaotus on sama igal katsel.

4.1 Paarikaupa varjatud tunnusega Markovi ahel

Vaatleme paarikaupa varjatud tunnusega Markovi ahelat. Üleminekumaatriks on genereeritud Dirichlet' jaotusest parameetritega $1, \ldots, 1$ ehk iga u_{t-1}, x_{t-1} korral on elemendivi-

isiliselt jaotus kirjeldatav kui

$$p(u_t, x_t|u_{t-1}, x_{t-1}) \sim Cat(A_{u_{t-1},x_{t-1}}), \qquad A_{u_{t-1},x_{t-1}} \sim Dir(1, \ldots, 1) \tag{4.1}$$

$$p(u_1, x_1) \sim Cat(A_1), \qquad A_1 \sim Dir(1, \ldots, 1). \tag{4.2}$$

Emissiooni y_t tiheduse määrab normaaljaotus keskväärtusega x_t ning standardhälbega 1.25 nagu ka (Soop, 2023) eksperimentides, ehk

$$Y_t|x_t \sim \mathcal{N}(y_t - x_t, 1.25). \tag{4.3}$$

Sellise mudeli puhul on võrdluseks kasutatud ka naiivset algoritmi, mis seab $x_t = \arg\min|x_t - y_t|$.

4.1.1 Eksperimendid väikese T korral

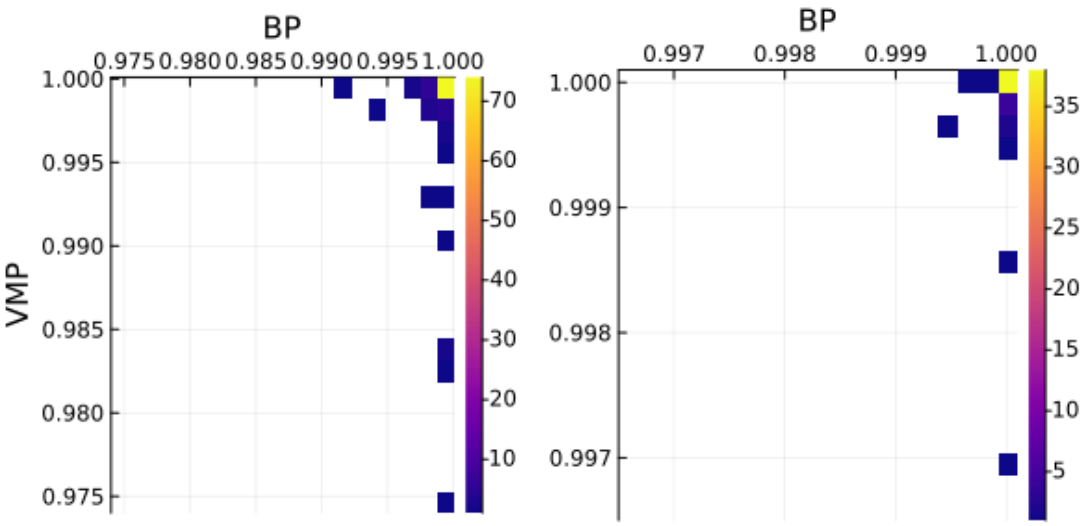
Uuritava mudeli korral paistavad kaks algoritmi üksteist komplementeerivat: $T = 20, N = 50$ korral oli näha, et 25 juhul saavutab BP parema raja ning 16 juhul VMP; $T = 15, N = 100$ korral oli BP algoritm 43 juhul parem, VMP 35 juhul parem.

Rohkem võrdlusi on näha joonisel 1 ning tabelis 1.

4.2 Vahelduva režiimiga mudel

Vaatleme vahelduva režiimiga mudelit, mida tutvustasime peatükis 1.1.1. Võtku U_t väärtusi hulgal $\{A, B, C\}$ ning X_t väärtusi hulgal $\{1, 2\}$. Olgu $\boldsymbol{y}$ vaatlusandmestik, mis on saadud $\{X_t\}$ müraga lugemisest, kus $Y_t = x_t + \varepsilon$, kus $\varepsilon = \mathcal{N}(0, 1.25)$.

Keskendume järgmisena mõõtmisviisidele, mille abil kirjeldada kolmekordset Markovi ahelat genereerivaid mudeleid. Lühikeste realisatsioonide, nt $T \in \{10, 15, 20\}$, korral me saame välja arvutada vastastikuse informatsiooni $D_{\mathrm{KL}}\left[p_x \| q_x\right]$, mis ei ole praktikas realistlik.



(a) $T = 15, N = 100$ (b) $T = 20, N = 50$

Joonis 1: VMP ja BP algortimide protsentiilide kahedimensionaalne histogramm N erineva katse mudeli (4.1), (4.2), (4.3) korral, mis on kirjeldatud peatükis 4.1. Leides kõikvõimalikud rajad $\{0,1\}^T$ saame BP ja VMP algoritmide abil Viterbi raja hinnangu $\boldsymbol{x}^*$ ja uurime, kui paljudest radadest on algoritmi hinnang parem osakaaluna — indikaatorfunktsiooni I abil kirjutame protsentiili kui $r(\boldsymbol{x^*}) = \sum_{\boldsymbol{x}} I_{p(\boldsymbol{x}^*)\geq p(\boldsymbol{x})}(\boldsymbol{x})/2^T$. Graafikul kujutame eksperimendi $n \in \{1, \ldots, N\}$ korral VMP ja BP radade protsentiilid $r_n^{\text{VMP}}, r_n^{\text{BP}}$ histogrammina. Kahe graafiku teljed pole võrdsed.

Tabel 1: Vaatleme mõlema algoritmi leitud Viterbi raja lähendit $\boldsymbol{x}^*$ N eksperimendi puhul kus kolmekaupa Markovi ahela realisatsiooni $(\boldsymbol{u}', \boldsymbol{x}', \boldsymbol{y})$ pikkus on T. Saame ka vaadelda eksperimentide puhul TMMi realisatsiooni jaoks ka originaalset rada $\boldsymbol{x}'$. Esimeses kolmes reas kujutame nende radade tõenäosusi protsentiilidena. Iga eksperimendi puhul vaatame iga algoritmi jaoks protsentiili kui $r(\boldsymbol{x^*}) = \sum_{\boldsymbol{x}} I_{p(\boldsymbol{x}^*) \geq p(\boldsymbol{x})}(\boldsymbol{x})/2^T$. Naiivse algoritmi puhul on Viterbi raja lähend iga t korral $x_t = \arg\min |x_t - y_t|$. Järgmises kolmes reas vaatleme vastavalt VMP, BP, naiivse algoritmi Viterbi raja lähendite $\boldsymbol{x}^*_{\text{VMP}}, \boldsymbol{x}^*_{\text{BP}}, \boldsymbol{x}^*_{\text{naive}}$ Hammingu kauguseid õigest Viterbi rajast $\boldsymbol{x}^*$. Viimases reas on originaalse raja Hammingu kaugus Viterbi rajast.

(a) $T = 15, N = 100$

	mean	min	max
$r(\boldsymbol{x}^*_{\text{BP}})$	0.9995	0.9919	1.0
$r(\boldsymbol{x}^*_{\text{VMP}})$	0.9986	0.9746	1.0
$r(\boldsymbol{x}')$	0.8851	0.09	0.9999
H vmp	1.63	0.0	7.0
H bp	1.61	0.0	7.0
H naive	2.68	0.0	8.0
H orig	5.22	1.0	12.0

(b) $T = 20, N = 50$

	mean	min	max
BP	0.99997	0.99955	1.0
VMP	0.99986	0.99701	1.0
ORIG	0.92382	0.48647	0.9998
H vmp	2.1	0.0	7.0
H bp	1.74	0.0	5.0
H naive	3.06	0.0	8.0
H orig	6.96	3.0	14.0

Kirjeldamaks, milline on fikseeritud vaatlusandmete korral mittehomogeenne paarikaupa Markovi ahel, vaatleme järgnevaid entroopiaid ja KL kaugusi:

$$\Delta_1 := \frac{1}{T} \sum_{t=1}^{T} D_{\text{KL}}[p_t \| p_{ut} \times p_{xt}]$$

$$\Delta_2 := \sum_{t=2}^{T} D_{\text{KL}}[p_{t-1,t} \| p_t \times p_{t-1}]$$

$$\Delta_3 := D_{\text{KL}}[p \| p_1 \times \ldots \times p_T]$$

$$\Delta_4 := H[q_x],$$

kus p_t on marginaali (U_t, X_t) jaotus $\sum_{\boldsymbol{u}_{\setminus t}, \boldsymbol{x}_{\setminus t}} = p(\boldsymbol{u}, \boldsymbol{x})$ ja $p_{u,t}(u_t) = \sum_{x_t} p_t(u_t, x_t)$, $p_{x,t}(x_t) = \sum_{u_t} p_t(u_t, x_t)$.

Me kasutame KL kauguste aritmeetilist keskmist Δ_1 mõõtmaks, kui lähedal on ka-

hekordse mittehomogeense Markovi ahela jaotus p jaotusele, mis võiks sobida BP algoritmi rakendamiseks hästi, suurused Δ_2, Δ_3 näitavad, kui lähedal on jaotus p jaotusele, mis sobib VMP rakendamiseks hästi. Kuigi oleks eelistatav suuruse Δ_1 asemel mõõta $D_{\mathrm{KL}}[p \| p_u \times p_x]$, siis see ei ole arvutuslikult realistlik ning loodame, et kui Δ_1 on väike, on ka $D_{\mathrm{KL}}[p \| p_u \times p_x]$ väike, mis peaks olema soodne tingimus BP algoritmile leidmaks Viterbi rada.

Entroopiat Δ_4 kasutame hindamaks, kui enesekindlad me võime olla selles, et Viterbi rada on leitud — maksimaalse entroopiaga jaotused annavad samuti BP ja VMP algoritmide tulemustena mingid Viterbi raja hinnangud, aga nende puhul on Viterbi raja tõenäosus väike. Kui q_x entroopia on väike, siis leidub selline $\boldsymbol{x}$, mille korral on $\sum_{\boldsymbol{u}} q_u(\boldsymbol{u}) \ln p(\boldsymbol{u}, \boldsymbol{x})$ keskmiselt suurem. Kasutame eksperimente uurimaks, kas võime sellest järeldada, et $p_x(\boldsymbol{x})$ on ka keskmiselt suurem teistest radadest.

4.2.1 Eksperimendid väikese T korral

Vaatleme $T \in \{10, 20\}$ korral vastavalt $N = 100$ ja $N = 50$ erinevat vahelduva režiimiga mudelit, kus me nõuame, et režiimisiseste üleminekumaatriksite p_A, p_B, p_C omavahelised KL kaugused oleksid suuremad kui 0.25. Igale mudelile genereerime ühe realisatsiooni.

Sobitame lineaarregressioonmudeli leidmaks vastavalt seoseid

- kas väiksem ühisinformatsioon Δ_1 ja väiksem entroopia Δ_4 ennustavad suuremat BP algoritmi Viterbi raja lähendi protsentiili ning
- kas väiksemad ühisinformatsioonid Δ_2, Δ_3 ja väiksem entroopia Δ_5 ennustavad suuremat VMP algoritmi Viterbi raja lähendi protsentiili.

Lineaarregressioon oodatud tulemusi ei andnud — oodates negatiivseid koeffitsente iga muutuja jaoks, olid võimalikust kümnest näitajast ülemine 95% usaldusintervall negatiivne vaid $\Delta_1^{T=10}$ ning $\Delta_5^{T=20}$ korral.

Tasub ka uurida, kas sellised mudelid ja realisatsioonid, kus BP või VMP algoritm leidis õige Viterbi raja ülesse, erinesid mingi Δ korral. Osutus, et Δ_3 on parim indikaator ennustamaks, kas on võimalik VMP algoritmiga jõuda Viterbi rajani — $\Delta_3^{T=10}$ oli keskeltläbi 6.4 korda väiksem skaleeritul kujul ja $\Delta_3^{T=20}$ oli 1.3 korda väiksem.

Tabel 2: Esimeses kahes reas kujutame leitud Viterbi raja lähendite protsentiile, kolmandas reas realisatsiooni $(\boldsymbol{u}', \boldsymbol{x}', \boldsymbol{y})$ raja $\boldsymbol{x}'$ protsentiili $r(\boldsymbol{x}')$ ning viimases neljas reas vastavalt VMP, BP, naiivse algoritmi lähendite ning originaalse raja $\boldsymbol{x}'$ Hammingu kauguseid Viterbi rajast.

(a) $T = 10, N = 100$

	mean	min	max
$r(\boldsymbol{x}^*_{\mathrm{BP}})$	0.9319	0.0586	1.0
$r(\boldsymbol{x}^*_{\mathrm{VMP}})$	0.9132	0.0264	1.0
$r(\boldsymbol{x}')$	0.8416	0.0957	1.0
H vmp	2.24	0.0	9.0
H bp	2.32	0.0	10.0
H orig	3.07	0.0	8.0
H naive	2.47	0.0	6.0

(b) $T = 20, N = 50$

	mean	min	max
$r(\boldsymbol{x}^*_{\mathrm{BP}})$	0.9731	0.6321	1.0
$r(\boldsymbol{x}^*_{\mathrm{VMP}})$	0.9451	0.1904	1.0
$r(\boldsymbol{x}')$	0.909	0.0007	1.0
H vmp	4.72	0.0	15.0
H bp	5.44	0.0	15.0
H orig	5.96	0.0	13.0
H naive	4.64	2.0	12.0

Võimalikke põhjusi, miks on raske ette ennustada, milline algoritm võiks paremini töötada, on mitmeid. Algoritm ise eeldab mitut õnnelikku kokkusattumust:

1. et KL kaugus $D_{\mathrm{KL}}[q_u \times q_x \| p]$ on hea lähend KL kaugusele $D_{\mathrm{KL}}[p \| q_u \times q_x]$,
2. et lähend q_x kirjeldab hästi marginaalprotsessi p_x,
3. ning, et q_x Viterbi rada on hea lähend marginaalprotsessi p_x Viterbi rajale.

Nüüd, kui prognoosida, kumb algoritm võiks mingi mudeli ja realisatsiooni korral paremini töötada, teeme veel lisaks järgmised eeldused:

4. marginaaljaotused $p_{u,t}, p_{u,x}$ on head lähendid BP algoritmi vahetulemusega saadud jaotustele q_u, q_x
5. mõõdul q_x on väike entroopia, kui mõõdul $\sum_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{x})$ on väike entroopia

6. marginaaljaotused $p_{t-1,t}, p_t, p_{t-1}$ on head lähendid jaotuste p ja $\prod_{t=1}^{T} q_t$ võrdlemiseks.

Eksperimentides loodud mudelite ja realisatsioonide korral osutus, et selliseid õnnelikke kokkusattumusi meil ei ole ja on raske ennustada, milline algoritm leiab suurema tõenäosusega Viterbi raja. Kuid hea oleks just selle teemaga tulevikus rohkem süvitsi minna.

4.2.2 Eksperimendid sama PMMi korral

Vaatleme järgmisena vahelduva režiimiga mudelit, kus paarikaupa Markovi protsessi jaotus on sama üle kõikide realisatsioonide. Vaatleme $N = 100$ ning $T = 15$ korral, kuidas käituvad erinevate realisatsioonide ja vaatluste $\boldsymbol{y}$ korral BP ja VMP algoritmid.

See osutus näiteks mudelist, kus genereeriti paarikaupa Markovi jaotus, kus suurema osa realisatsioonide korral on VMP algoritm parem. Esitame taas tulemused histogrammina joonisel 2 ja tabelis 3, kus seekord tunneme huvi ka sellest, et millised on KL kaugused marginaaljaotusest p_x ning kui kaugel on esimesed viis kõige tõenäolisemat rada üksteisest Hammingu kaugu mõttes — täpsemalt uurime 2., 3., 4. ja 5. raja Hammingu kaugusi Viterbi rajast.

Katsetel oli 66 juhul sajast

$$D_{\mathrm{KL}}\left[p_x \middle\| \sum_{\boldsymbol{u}} \prod_{t=1}^{T} q_t\right] < D_{\mathrm{KL}}\left[p \| q_x\right],$$

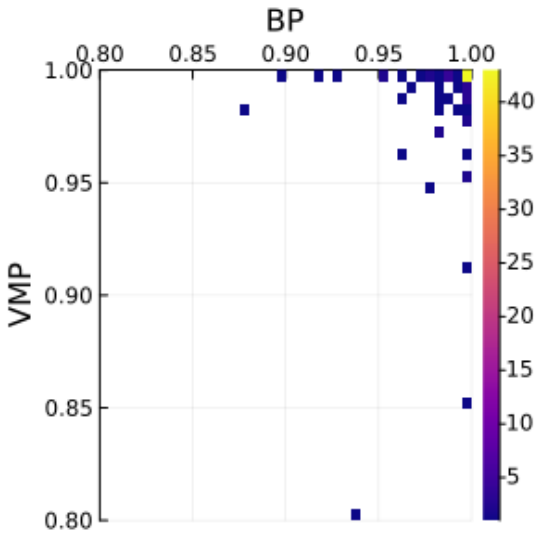
ning ka 85 juhul sajast

$$D_{\mathrm{KL}}\left[\sum_{\boldsymbol{u}} \prod_{t=1}^{T} q_t \middle\| p_x\right] < D_{\mathrm{KL}}\left[q_x \| p_x\right],$$

mis kinnitab nõrgalt, et leidub mudeleid, mille korral on nii

$$D_{\mathrm{KL}}\left[\sum_{\boldsymbol{u}} \prod_{t=1}^{T} q_t \middle\| p_x\right], D_{\mathrm{KL}}\left[p_x \middle\| \sum_{\boldsymbol{u}} \prod_{t=1}^{T} q_t\right]$$

(need ei ole suure T korral leitavad), kui ka $D_{\mathrm{KL}}[\prod_{t=1}^{T} q_t \| p]$ väikesed ning tõenäolisemalt



Joonis 2: VMP ja BP algortimide protsentiilide kahedimensionaalne histogramm $N = 100$ erineva katse korral, kus $T = 15$ ja paarikaupa Markovi ahel on igal katsel sama. Graafikul kujutame eksperimendi $n \in \{1, \ldots, N\}$ korral VMP ja BP algoritmide Viterbi radade lähendite protsentiilid $r(\boldsymbol{x}^*_{\mathrm{VMP}}), r(\boldsymbol{x}^*_{\mathrm{BP}})$ histogrammina.

ka VMP algoritmi Viterbi hinnang parem BP algoritmist.

Tabel 3: VMP ja BP algortimide tulemused $N = 100$ erineva katse korral, kus $T = 15$ ja paarikaupa Markovi ahel on igal katsel sama. Leiame protsentiilid $r(\boldsymbol{x}^*_{\mathrm{VMP}}), r(\boldsymbol{x}^*_{\mathrm{BP}})$.
Järgmisena leiame KL kaugused marginaaljaotuse p_x ja lähendite vahel.
Leiame absoluutarvu radadest, mille marginaaltõenäosus oli suurem hinnangutest.
Leiame Hammingu kaugused erinevate algoritmide puhul ning viimasel neljal real kujutame 2., 3., 4. ja 5. kõige tõenäolisema raja Hammingu kaugusi Viterbi rajast.

	mean	min	max
$r(\boldsymbol{x}^*_{\mathrm{BP}})$	0.9403	0.005	1.0
$r(\boldsymbol{x}^*_{\mathrm{VMP}})$	0.9671	0.5037	1.0
$D_{\mathrm{KL}}\left[p_x \| q_x\right]$	11.5	0.7671	62.78
$D_{\mathrm{KL}}\left[p_x \| \sum_{\boldsymbol{u}} \prod_{t=1}^{T} q_t\right]$	5.145	0.5632	22.25
$D_{\mathrm{KL}}\left[q_x \| p_x\right]$	7.313	0.6023	25.4
$D_{\mathrm{KL}}\left[\sum_{\boldsymbol{u}} \prod_{t=1}^{T} q_t \| p_x\right]$	2.67	0.3981	7.225
$\sum_{\boldsymbol{x}} I_{p(\boldsymbol{x})>p(\boldsymbol{x}^*_{\mathrm{BP}})}(\boldsymbol{x})$	1955.0	0.0	32600.0
$\sum_{\boldsymbol{x}} I_{p(\boldsymbol{x})>p(\boldsymbol{x}^*_{\mathrm{VMP}})}(\boldsymbol{x})$	1955.0	0.0	32600.0
H vmp	3.24	0.0	12.0
H bp	3.74	0.0	14.0
H orig	4.44	0.0	10.0
H naive	3.06	0.0	7.0
H_2	1.36	1.0	6.0
H_3	1.4	1.0	5.0
H_4	1.75	1.0	7.0
H_5	1.69	1.0	5.0

Kokkuvõte

Tutvusime Viterbi raja

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathcal{X}^T} \sum_{\boldsymbol{u}\in\mathcal{U}^T} p(\boldsymbol{x}, \boldsymbol{u})$$

lähendamise probleemiga juhul, kus p on kolmekaupa Markovi ahela jaotus ning vaatlused on fikseeritud. Lähendite leidmise kaks algoritmi põhinevad variatsioonilistel meetoditel. Kirjeldati *belief propagation* (BP) ja *variational message passing* (VMP) algoritme, mille puhul erineb optimiseeritava jaotuste hulk, ehk mis kitsendusi nõutakse — BP algoritmi puhul $q(\boldsymbol{u}, \boldsymbol{x}) = q(\boldsymbol{u})q(\boldsymbol{x})$ ning VMP algoritmi puhul $q(\boldsymbol{u}, \boldsymbol{x}) = \prod_{t=1}^{T} q_t(u_t, x_t)$. Tõestasime, et kui me minimiseerime KL kaugust $D_{\mathrm{KL}}[q\|p]$, kus q on optimiseeritav jaotus kitsendusega, siis mõlemad algoritmid leiavad ühese lahendi q_∞. Kusjuures nende lahendite Viterbi rada on arvutuslikult leida palju lihtsam.

Implementeeriti BP ja VMP algoritmid, mille abil leida kolmekaupa Markovi ahela Viterbi raja kaks lähendit. Kuigi BP algoritm on arvutuslikult keerukam, leidus eksperimente, kus paremad oli VMP algoritmi lähendid. Erinevatel katsetel oli ühe algoritmi Viterbi raja lähend õige Viterbi rada, kuid teise algoritmi lähend mitte. Osutus keeruliseks leida häid meetrikaid kirjeldamaks jaotuse ja vaatluste $\boldsymbol{y}$ korral, millist algoritmi rakendada parema lähendi saamiseks.

Seepärast oleks kasulik edaspidi põhjalikumalt uurida, kas erinevatel tingimustel on siiski võimalik efektiivselt klassifitseerida jaotusi prognoosimaks, milline lähendusalgoritm võiks parim olla. Seda võiks teha lisades veel võrdlusesse teisi lähendusalgoritme ning rohkemaid kolmekaupa Markovi mudeleid kui ainult vahelduva režiimiga mudel.

Väärib mainimist, et leides kahe algoritmi korral lähendjaotused q, võib neid jaotusi kasutada ka teiste ülesannete korral, kus Viterbi raja leidmine ei ole põhifookus.

Kasutatud allikad

Avans, K. (2021). *Paarikaupa Markovi mudel: definitsioon ja näited*. Tartu Ülikooli magistritöö.

Bagaev, Dmitry, Albert Podusenko **and** Bert de Vries (2023). "RxInfer: A Julia package for reactive real-time Bayesian inference". **in***Journal of Open Source Software*: 8.84, **page** 5161. DOI: 10.21105/joss.05161. URL: https://doi.org/10.21105/joss.05161.

Kuljus, Kristi **and** Jüri Lember (2022). *Hybrid classifiers of pairwise Markov models*. arXiv: 2203.10574 [stat.ME]. URL: https://arxiv.org/abs/2203.10574.

Lember, J. (2022). *Informatsiooniteooria, loengukonspekt ja ülesanded*. URL: https://courses.ms.ut.ee/MTMS.02.040/2022_spring/uploads/Main/info22.pdf.

Log sum inequality (2025). *Log sum inequality — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-August-2025]. URL: https://en.wikipedia.org/wiki/Log_sum_inequality.

Lyngsø, Rune B. **and** Christian N.S. Pedersen (2002). "The consensus string problem and the complexity of comparing hidden Markov models". **in***Journal of Computer and System Sciences*: 65.3. Special Issue on Computational Biology 2002, **pages** 545–569. ISSN: 0022-0000. DOI: https://doi.org/10.1016/S0022-0000(02)00009-0. URL: https://www.sciencedirect.com/science/article/pii/S0022000002000090.

Oja, E. **and** P. Oja (1991). *Funktsionaalanalüüs*. Tartu Ülikool.

Parr, T., D. Markovic, S.J. Kiebel **and** K.J. Friston (2019). "Neuronal message passing using Mean-field, Bethe, and Marginal approximations". **in***Sci Rep 9, 1889*.

Pihel, J.E. (2024). *VBMC.jl*. https://github.com/jaanerik/VBMC.jl.

Soop, O (2023). *Kolmekaupa Markovi ahelate Viterbi raja lähendamine*. Tartu Ülikooli magistritöö.

Lisa 1. Julia koodi repositoorium

Koodirepositoorium, mida kasutati algoritmide implementeerimiseks ning eksperimentatsioonideks, on leitav GitHubis (Pihel, 2024). Kui BP algoritm on implementeeritud koos vajalike jaotuste ja struktuuridega ning on põhjus, miks repositooriumi ülesehitus on omane Julia teegile, siis algseks inspiratsiooniks valida just selline ebatavaline programmeerimiskeel oli teek RxInfer (Bagaev, Podusenko **and** Vries, 2023), mis on lihtsustanud paljude *variational inference* meetodite implementatsiooni ning võimaldas VMP algoritmi kirjeldada palju lühemalt.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Jaan Erik Pihel,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Kolmekaupa Markovi ahelate Viterbi raja lähendamine variatsiooniliste meetoditega, mille juhendajad on Jüri Lember ja Oskar Soop, 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.

Jaan Erik Pihel
20.08.2025