УДК 519.217

ОЦІНЮВАННЯ ХАРАКТЕРИСТИК ЧАСТКОВО СПОСТЕРЕЖУВАНОГО ЛАНЦЮГА МАРКОВА НА ДВІЙКОВИХ ПОСЛІДОВНОСТЯХ

А. В. Цибульник[1,a], І. І. Ніщенко[1]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

Об'єктом дослідження є ланцюг Маркова зі значеннями в множині двійкових послідовностей фіксованої довжини. Динаміка ланцюга задається як випадкове блукання вершинами одиничного куба, розмірність якого збігається з довжиною двійкової послідовності. Стани цього ланцюга є неспостережуваними (прихованими), а матриця перехідних імовірностей — невідомою.

Спостережуваними величинами в кожен момент часу є набір значень певного функціонала від фіксованих підмножин двійкової послідовності, яка описує поточний стан прихованого ланцюга. Також є відомими значення вказаного функціонала, обчисленого від деякої невідомої підмножини стану прихованого ланцюга. Задача полягає у локалізації — оцінюванні потужності та набору елементів цієї підмножини. Для розв'язування задачі використовується математичний апарат прихованих марковських моделей.

Ключові слова: ланцюг Маркова, модель Еренфестів, алгоритм Баума-Велша, алгоритм Вітербі

Вступ

Марковські моделі мають широкий та ефективний арсенал інструментів для аналізу динаміки систем, поведінка яких у кожен наступний момент часу зумовлюється лише поточним станом системи та не залежить від характеру еволюції у попередні моменти часу.

Наприклад, в біоінформатиці [1, глава 9] апарат ланцюгів Маркова застосовують при дослідженні еволюції молекул ДНК протягом певного часу, вважаючи при цьому за стан системи зв'язану послідовність так званих нуклеотидів, які формуються над алфавітом азотистих основ $\{T, C, A, G\}$.

Водночас, у випадку, коли безпосереднє спостереження еволюції ланцюга Маркова є неможливим чи обмеженим, застосовують моделі прихованих марковських ланцюгів. У такому випадку аналіз поведінки процесу відбувається за деякою опосередкованою інформацією про «приховані», справжні стани ланцюга.

Вважаючи, що динаміка ланцюга відбувається згідно узагальненої моделі Еренфестів, у цій роботі було застосовано приховану марковську модель для аналізу еволюції послідовностей, побудованих над алфавітом бінарних символів $\{0, 1\}$.

1. Моделювання об'єкту дослідження

Розглянемо ланцюг Маркова $\left\{X^t\right\}_{t=\overline{1,T}}$, який приймає значення зі скінченної множини $E = \{0,1\}^N$ — множини всеможливих бінарних послідовностей довжини N.

[a]anton.tsybulnik@gmail.com

Динаміка ланцюга відбувається згідно узагальненої моделі Еренфестів: в кожен момент часу t навмання обирається число j з множини індексів $\{1, 2, \dots, N\}$ бінарної послідовності X^t та відповідний елемент стану X^t_j залишається незмінним з імовірністю p або змінюється на протилежний бінарний символ з імовірністю $1 - p$.

Як наслідок окресленої динаміки, матриця перехідних імовірностей ланцюга матиме вигляд:

$$A_{xx'} = P\left(X^{t+1} = x' \mid X^t = x\right) = \begin{cases} p, & x' = x \\ \dfrac{1-p}{N}, & x'_j = 1 - x_j \\ & \forall i \neq j : x'_i = x_i \\ 0, & \text{інакше} \end{cases}$$

Крім того, інваріантний розподіл $\pi = \left(\pi_x\right)_{x\in E}$ заданого ланцюга є рівномірним, тобто $\pi_x = \frac{1}{2^N}$. Вважатимемо, що початковий розподіл збігається з π.

Наступним кроком введемо послідовність випадкових величин $\left\{Y^t\right\}_{t=\overline{1,T}}$, які формуються таким чином:

$$Y^t = \left(Y^t_k\right)_{k=\overline{1,L}} = \left(\phi\left(X^t, I_k\right)\right)_{k=\overline{1,L}},\ t = \overline{1,T}, \qquad (1.1)$$

де $I = \left\{I_1, \dots, I_L\right\}$ — задані підмножини множини індексів $\{1, 2, \dots, N\}$, а функціонал ϕ визначимо так:

$$\phi\left(X^t, I_k\right) = \sum_{i\in I_k} X^t_i \qquad (1.2)$$

Твердження 1. *Послідовність* $\left\{\left(X^t, Y^t\right)\right\}_{t=\overline{1,T}}$ *утворює приховану марковську модель* (π, A, B)*, де*

$$B_{xy} = P\left(Y^t = y \mid X^t = x\right) = \prod_{k=1}^{L} \mathbb{1}\left(y_k = \sum_{i\in I_k} x_i\right)$$

2. Постановка задачі

За спостереженнями (1.1) прихованої марковської моделі слід знайти розв'язки задач:

1. Оцінити невідомий «параметр мутації» p елементів бінарних послідовностей прихованого ланцюга Маркова та декодувати послідовність станів прихованого ланцюга;
2. Вважаючи, що спостерігається деяке додаткове значення функціонала (1.2) від прихованих станів ланцюга по невідомій «множині неявних індексів» I_*, оцінити потужність цієї множини та відтворити набір її елементів;
3. Вважаючи, що значення введеного функціонала (1.2) від прихованих станів ланцюга по множинам $I_1, \dots, I_L$ спостерігаються так:

$$\phi\left(X^t, I_k\right) = \sum_{i \in I_k} \widetilde{X}_i^t, \ k = \overline{1, L}, \tag{2.1}$$

де для $i \in I_k$

$$\widetilde{X}_i^t = \begin{cases} 1 - X_i^t, & \text{з імовірністю } q_k \\ X_i^t, & \text{з імовірністю } 1 - q_k \end{cases}, \tag{2.2}$$

оцінити невідомий параметр моделі p та ймовірності спотворень $q_1, q_2, \dots, q_L$.

2.1. Оцінка невідомого параметра моделі

Алгоритм навчання Баума-Велша

Спостерігаючи (1.1), скористаємося методом максимальної правдоподібності, шукаючи оцінку невідомого параметра p таким чином:

$$\widehat{p} = \underset{p}{\operatorname{argmax}} \sum_{x \in E^T} L_{p,x,y},$$

де

$$L_{p,x,y} = P\left(X = x, Y = y \mid p\right) \tag{2.3}$$

$$X = x \Longleftrightarrow \left(X^1 = x^1, \dots, X^T = x^T\right)$$

$$Y = y \Longleftrightarrow \left(Y^1 = y^1, \dots, Y^T = y^T\right)$$

Щоправда, для заданої марковської моделі вигляд функції правдоподібності (2.3) матиме громіздкий та неможливий для безпосереднього диференціювання вигляд.

Однак, в такому випадку можна застосувати модифікацію ЕМ-алгоритму для дослідження прихованих ланцюгів Маркова — ітераційний алгоритм Баума-Велша [1, розділ 15].

Задавши деяке наближення $p^{(0)}$ невідомого параметра p, покладемо

$$p^{(n+1)} = \underset{p}{\operatorname{argmax}}\, Q\left(p^{(n)}, p\right),$$

де

$$Q\left(p^{(n)}, p\right) = \sum_{x \in E^T} L_{p^{(n)},x,y} \cdot \ln L_{p,x,y} \tag{2.4}$$

є так званою функцією квазі-log правдоподібності.

Доведено [1, розділ 4], що така ітераційна процедура є збіжною і приводить до точки локального максимуму логарифму функції правдоподібності (2.3).

Максимізація функції (2.4) приводить до такої ітераційної формули переоцінки параметра p :

$$p^{(n+1)} = p^{(n)} \cdot \frac{\sum\limits_{t=1}^{T-1} \sum\limits_{x \in E} \alpha_t(x)\, B_{xy^{t+1}}\, \beta_{t+1}(x)}{\sum\limits_{t=1}^{T-1} \sum\limits_{x \in E} \alpha_t(x)\, \beta_t(x)}, \tag{2.5}$$

де

$$\alpha_t(x) = P\left(Y^1 = y^1, \dots, Y^t = y^t,\ X^t = x \mid p^{(n)}\right) \tag{2.6}$$

$$\beta_t(x) = P\left(Y^{t+1} = y^{t+1}, \dots, Y^T = y^T \mid X^t = x,\ p^{(n)}\right) \tag{2.7}$$

є так званими коефіцієнтами прямого та зворотного ходу відповідно [2, розділ 5].

Алгоритм декодування Вітербі

Використовуючи оцінене значення параметра $\widehat{p}$, отримане в результаті застосування алгоритму навчання Баума-Велша, скористаємося алгоритмом декодування Вітербі [2, розділ 6] для пошуку такої послідовності прихованих станів $\widehat{X}^1, \widehat{X}^2, \dots, \widehat{X}^T$, яка найкращим чином описує наявні спостереження:

$$\widehat{X} = \underset{x \in E^T}{\operatorname{argmax}}\, P\left(X = x \mid Y = y, \widehat{p}\right)$$

2.2. Оцінка множини неявних індексів

Нехай окрім набору спостережень (1.1) протягом еволюції ланцюга на кожному кроці t спостерігається деяке додаткове значення $Y_{I_*}^t$ функціонала (1.2) від прихованого стану ланцюга по деякій невідомій підмножині індексів $I_* \subseteq \{1, 2, \dots, N\}$:

$$Y_{I_*} = \left(Y_{I_*}^t\right)_{t=\overline{1,T}} = \left(\sum_{i \in I_*} X_i^t\right)_{t=\overline{1,T}}$$

Перш за все, оцінимо потужність множини I_*. Зауважимо, що в силу заданого способу еволюції прихованого ланцюга Маркова

$$P\left(Y_{I_*}^t = Y_{I_*}^{t+1}\right) = \frac{\left|I_*\right|}{N} \cdot p + \frac{N - \left|I_*\right|}{N}$$

Ця рівність дозволяє побудувати незміщену та змістовну оцінку для потужності $|I_*|$.

Твердження 2. *Змістовною і незміщеною оцінкою потужності множини I_* є статистика*

$$\widehat{|I_*|} = \frac{N}{1-p}\left(1 - \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{1}\left(Y_{I_*}^t = Y_{I_*}^{t+1}\right)\right) \tag{2.8}$$

Аналогічним чином побудуємо оцінку для потужності перетину множини I_* з індексами множин, які задають спостереження моделі. Вказана оцінка дозволить виявити взаємне розташування елементів множини неявних індексів та множини доступних для дослідження елементів прихованого стану ланцюга Маркова.

Твердження 3. *Нехай $H \subseteq I_1 \cup I_2 \cup \ldots \cup I_L$ — довільна підмножина множини спостережуваних індексів $I_1 \cup I_2 \cup \ldots \cup I_L$. Тоді змістовною та незміщеною оцінкою потужності множини $I_* \cap H$ є статистика*

$$|\widehat{I_* \cap H}| = \frac{N}{(T-1)(1-p)} \cdot \sum_{t=1}^{T-1} \mathbb{1}\left(Y^t_{I_*} \neq Y^{t+1}_{I_*},\, Y^t_H \neq Y^{t+1}_H\right) \tag{2.9}$$

Стратегія визначення елементів, які безпосередньо входять в множину I_*, складатиметься з декількох кроків:

- із загальної множини індексів $\{1, 2, \ldots, N\}$ сформувати всеможливі підмножини довжиною $\widehat{|I_*|}$, тобто вибірку

$$\left\{\mathtt{I}_1, \mathtt{I}_2, \ldots, \mathtt{I}_{C_N^{\widehat{|I_*|}}}\right\} \tag{2.10}$$

- для кожного «кандидата» $\mathtt{I}_k$ з множини (2.10) згенерувати послідовність значень функціонала (1.2) від декодованих прихованих станів по відповідних індексах:

$$\widehat{Y}_{\mathtt{I}_k} = \left(\widehat{Y}^t_{\mathtt{I}_k}\right)_{t=\overline{1,T}} = \left(\sum_{i \in \mathtt{I}_k} \widehat{X}^t_i\right)_{t=\overline{1,T}}$$

- за допомогою деякої заданої міри d оцінити для кожного $\mathtt{I}_k$ відстань між наборами $\widehat{Y}_{\mathtt{I}_k}$ та Y_{I_*};
- оцінкою $\widehat{I}$ множини I_* стане той «кандидат» $\mathtt{I}_k$ з множини (2.10), для якого d буде найменшою:

$$\widehat{I} = \underset{1 \leqslant k \leqslant C_N^{\widehat{|I_*|}}}{\operatorname{argmin}}\, d\left(\widehat{Y}_{\mathtt{I}_k}, Y_{I_*}\right) \tag{2.11}$$

Міру близькості d між двома невід'ємними цілочисельними множинами $\widehat{Y}_{\mathtt{I}_k}$ та Y_{I_*} однакової довжини визначатимемо або за допомогою середньоквадратичної відстані

$$d_S\left(\widehat{Y}_{\mathtt{I}_k}, Y_{I_*}\right) = \sum_{t=1}^{T}\left(\widehat{Y}^t_{\mathtt{I}_k} - Y^t_{I_*}\right)^2, \tag{2.12}$$

або користуючись зваженою відстанню Жаккара [3]

$$d_J\left(\widehat{Y}_{\mathtt{I}_k}, Y_{I_*}\right) = 1 - \frac{\sum_{t=1}^{T} \min\left(\widehat{Y}^t_{\mathtt{I}_k}, Y^t_{I_*}\right)}{\sum_{t=1}^{T} \max\left(\widehat{Y}^t_{\mathtt{I}_k}, Y^t_{I_*}\right)} \tag{2.13}$$

2.3. Оцінка коефіцієнтів спотворення

Припустимо, що значення функціонала (1.2) від прихованих станів ланцюга $\{X^t\}_{t=\overline{1,T}}$ по множинам $I_1, \ldots, I_L$ спостерігаються із деякими ймовірностями спотворення $q_1, q_2, \ldots, q_L$ згідно (2.1) та (2.2).

Оцінимо параметр p та вектор імовірностей спотворень $q = (q_1, q_2, \ldots, q_L)$, використовуючи ітераційний алгоритм Баума-Велша.

Твердження 4. *Якщо множини $I_1, \ldots, I_L$ є попарно неперетинними, то утворена послідовність $\left\{\left(X^t, Y^t\right)\right\}_{t=\overline{1,T}}$ є прихованою марковською моделлю $\left(\pi, A, B^q\right)$, де*

$$B^q_{xy} = P\left(Y^t = y \mid X^t = x\right) = \prod_{k=1}^{L} P\left(\xi^k_{01}(x) + \xi^k_{11}(x) = y_k\right),$$

і для довільного $k = \overline{1, L}$

$$\xi^k_{01}(x) \sim Bin\left(|I_k| - \sum_{i \in I_k} x_i,\, q_k\right)$$
$$\xi^k_{11}(x) \sim Bin\left(\sum_{i \in I_k} x_i,\, 1 - q_k\right)$$

є незалежними випадковими величинами.

Виберемо деяке початкове наближення моделі $\left(\pi, A^{(0)}, B^{q^{(0)}}\right)$, визначимо коефіцієнти прямого (2.6) та зворотного (2.7) ходу. Тоді ітераційна формула переоцінки параметра p матиме вид:

$$p^{(n+1)} = p^{(n)} \cdot \frac{\sum_{t=1}^{T-1} \sum_{x \in E} \alpha_t(x)\, B^{q^{(n)}}_{xy^{t+1}}\, \beta_{t+1}(x)}{\sum_{t=1}^{T-1} \sum_{x \in E} \alpha_t(x)\, \beta_t(x)}, \tag{2.14}$$

а формула переоцінки компонент вектора $\left(q_k\right)_{k=\overline{1,L}}$:

$$q_k^{(n+1)} = q_k^{(n)} \cdot \frac{\sum_{t=1}^{T} \sum_{x \in E} \beta_t(x) \sum_{x' \in E} \alpha_{t-1}(x')\, A^{(n)}_{x'x} \sum_{i \in I_k} P^{q^{(n)}}_{x,i}}{|I_k| \sum_{t=1}^{T} \sum_{x \in E} \alpha_t(x)\, \beta_t(x)}, \tag{2.15}$$

де при $i \in I_m$

$$P^q_{x,i} = P\left(\widetilde{\xi^m_{01}}(x) + \widetilde{\xi^m_{11}}(x) = y_m + x_i - 1\right) \times$$
$$\times \prod_{\substack{k=\overline{1,L} \\ k \neq m}} P\left(\xi^k_{01}(x) + \xi^k_{11}(x) = y_k\right)$$

та

$$\widetilde{\xi^m_{01}}(x) \sim Bin\left(|I_m| - 1 - \sum_{j \in I_m \setminus \{i\}} x_j,\, q_m\right)$$
$$\widetilde{\xi^m_{11}}(x) \sim Bin\left(\sum_{j \in I_m \setminus \{i\}} x_j,\, 1 - q_m\right)$$

Наостанок зауважимо, що при великих значеннях довжини ланцюга ($T > 300$) виникає потреба у шкалюванні [2, розділ 5] коефіцієнтів прямого та зворотного ходу, адже їхні значення стають нерозрізнювано малими для обчислювальних ресурсів. Процедура нормування не вносить змін у вигляд ітераційних формул переоцінки (2.5), (2.14) чи (2.15).

3. Результати чисельного експерименту

Оцінка невідомого параметра моделі

Було згенеровано прихований ланцюг Маркова протягом $T = 200$ моментів часу для бінарних послідовностей довжини $N = 5$ при заданому параметрі моделі $p = 0.2$. Множину спостережуваних індексів було задано таким чином:

$$I = \{I_1, I_2\} = \{(2,3), (1,4)\} \tag{3.1}$$

Рис. 1 демонструє збіжність алгоритму навчання Баума-Велша при оцінці параметра p. Червоним кольором позначено початкове наближення $p^{(0)} = 0.55$.

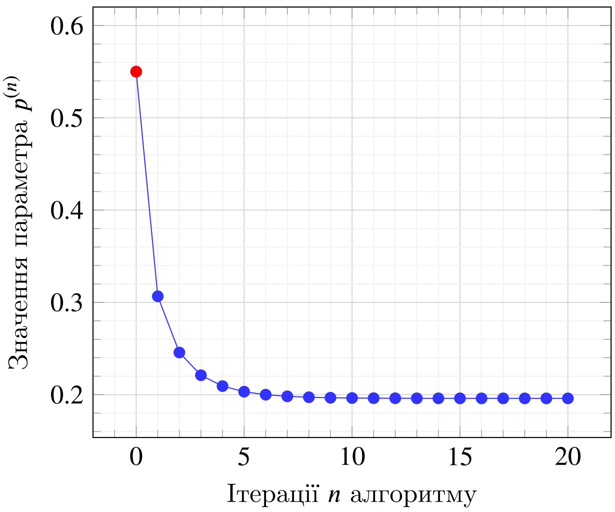

Рис. 1. Ітерації алгоритму Баума-Велша для оцінки параметра p

Вже за $n = 12$ ітерацій алгоритм досягає точності переоцінки $\varepsilon = 0.0001$ оцінюваного параметра. При цьому, отримане значення $\widehat{p} = 0.1959$ відрізняється від свого істинного значення $p = 0.2$ на $\delta = 0.0041$.

Алгоритм декодування прихованих станів

Наступним кроком, отримавши оцінене значення $\widehat{p}$, декодуємо ланцюг прихованих станів за допомогою алгоритму Вітербі [2, розділ 6].

Якість отриманих результатів оцінимо через порівняння в кожен момент часу t істинної прихованої бінарної послідовності X^t та декодованої $\widehat{X^t}$ за допомогою відстані Геммінга:

$$d_H\left(X^t, \widehat{X^t}\right) = \sum_{i=1}^{N} \mathbb{1}\left(X_i^t \neq \widehat{X_i^t}\right)$$

Таким чином, чим більше символів між справжнім та декодованим станами збігаються, тим меншою буде відповідна відстань Геммінга d_H. З гістограми результатів (Рис. 2) видно, що 17% усього ланцюга декодовано правильно.

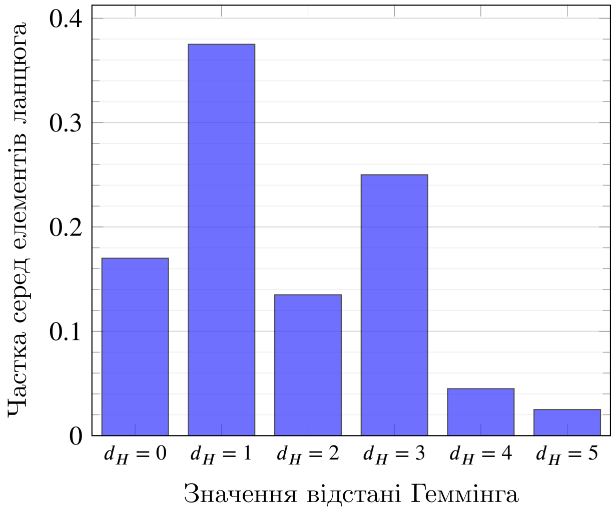

Рис. 2. Результати алгоритму декодування Вітербі

Наявність близько 40% помилок в одному символі може бути наслідком того, що одного елемента стану немає серед спостережуваних областей ланцюга. Крім того, оцінений параметр $\widehat{p}$ має похибку $\delta = 0.0041$ відносно свого істинного значення, що також впливає на результати задачі декодування.

Оцінка множини неявних індексів

В ролі множини неявних індексів було обрано набір $I_* = (1,3,5)$. В Табл. 1 показано збіжність змістовної та незміщеної оцінки (2.8) потужності $\widehat{|I_*|}$.

Таблиця 1. Оцінка потужності $\widehat{|I_*|}$ при збільшенні довжини ланцюга T

T	200	400	600	800	1000
$\widehat{p}$	0.1959	0.1823	0.1882	0.2099	0.2092
$\widehat{\lvert I_*\rvert}$	2	2	2	3	3

Бачимо, що довжини ланцюга $T = 200$ недостатньо для отримання точної оцінки. Однак, оскільки обране значення N є невеликим, для оцінки потужності множини неявних індексів в такому випадку можна використати емпіричну оцінку вигляду:

$$\widehat{|I_*|} = \max_{1 \leqslant t \leqslant T} Y_{I_*}^t$$

Застосуємо отримане значення потужності для виразу (2.11), щоб віднайти елементи, які безпосередньо входять в I_*: квадратична відстань (2.12) вказує на сукупність $\widehat{I}_S = (1,2,5)$, а зважена відстань Жаккара (2.13) — на сукупність $\widehat{I}_J = (1,2,3)$.

Дилему можна вирішити шляхом збільшення T та подальшого використання змістовної оцінки (2.9) для визначення взаємного розташування елементів множини неявних індексів відносно спостережуваних індексів (3.1).

Оцінка коефіцієнтів спотворення

Для кожної із спостережуваних областей (3.1) змодельованого ланцюга було обрано такі ймовірності викривлення: $q = (q_1, q_2) = (0.05, 0.1)$.

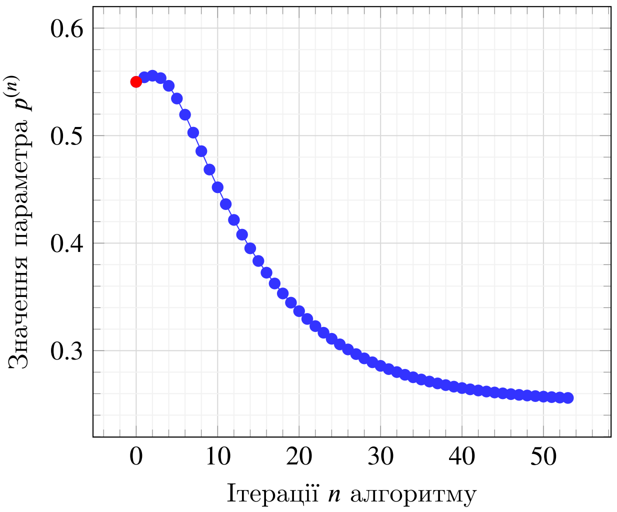

Рис. 3. Ітерації алгоритму Баума-Велша для оцінки параметра p, враховуючи спотворення спостережень

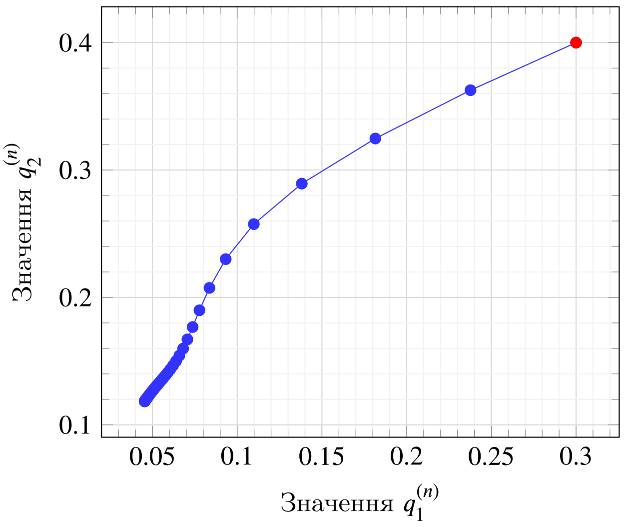

Рис. 4. Ітерації алгоритму Баума-Велша для оцінки компонент вектора q, враховуючи спотворення спостережень

Рис. 3 та Рис. 4 демонструють результати переоцінки невідомих параметрів моделі. Червоним кольором позначене початкове наближення $p^{(0)} = 0.55$ та $q^{(0)} = (0.3, 0.4)$.

Для досягнення аналогічної точності переоцінки $\varepsilon = 0.0001$ оцінюваного параметра p у випадку спотворенних даних знадобилося $n = 53$ ітерацій алгоритму. При цьому, помітне збільшення похибки: отримане значення $\widehat{p} = 0.2559$ відрізняється від свого істинного значення $p = 0.2$ на суттєво вищий показник $\delta = 0.0559$.

В той же час, точність оцінки коефіцієнтів спостворення $\widehat{q} = \left(\widehat{q}_1, \widehat{q}_2\right) = (0.0454, 0.1184)$ є високою: $\delta = (\delta_1, \delta_2) = (0.0046, 0.0184)$.

Висновки

В роботі було розглянуто задачу оцінювання певних характеристик ланцюга Маркова, змодельованого на бінарних послідовностях фіксованої довжини: невідомі параметри моделі були оцінені або шляхом побудови змістовних та незміщених статистичних оцінок, або за допомогою ітераційного алгоритму Баума-Велша.

Результати чисельного експерименту продемонстрували ефективність використаних методів, зокрема збіжність побудованих оцінок до істинних значень параметрів.

Перелік використаних джерел

1. *Koski T.* Hidden Markov models for bioinformatics. — 2002-е вид. — New York, NY : Springer, 11.2001. — (Computational Biology).
2. *Nilsson M.* First Order Hidden Markov Model: Theory and Implementation Issues : тех. звіт. / Blekinge Institute of Technology, School of Engineering, Department of Signal Processing. — 2005.
3. Finding the Jaccard Median / F. Chierichetti, R. Kumar, S. Pandey, S. Vassilvitskii // Proceedings of the 2010 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). — 2010. — С. 293—311. — DOI: 10.1137/1.9781611973075.25. — eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611973075.25. — URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611973075.25.