

# Оцінювання характеристик частково спостережуваного ланцюга Маркова на двійкових послідовностях

А. В. Цибульник    І. І. Ніщенко

Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених

2023

# План доповіді

- 1 Сфери застосування бінарних послідовностей
- 2 Моделювання об'єкту дослідження
- 3 Побудова оцінок невідомих параметрів
- 4 Результати чисельного експерименту

# План доповіді

- 1 Сфери застосування бінарних послідовностей
- 2 Моделювання об'єкту дослідження
- 3 Побудова оцінок невідомих параметрів
- 4 Результати чисельного експерименту

# Сфери застосування бінарних послідовностей

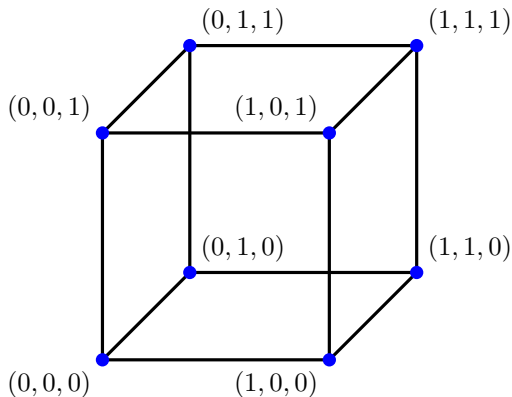
- Еволюція ДНК в біології
- Бінарні послідовності в теорії інформації
- Спінові системи у фізиці

# План доповіді

- 1 Сфери застосування бінарних послідовностей
- 2 Моделювання об'єкту дослідження**
- 3 Побудова оцінок невідомих параметрів
- 4 Результати чисельного експерименту

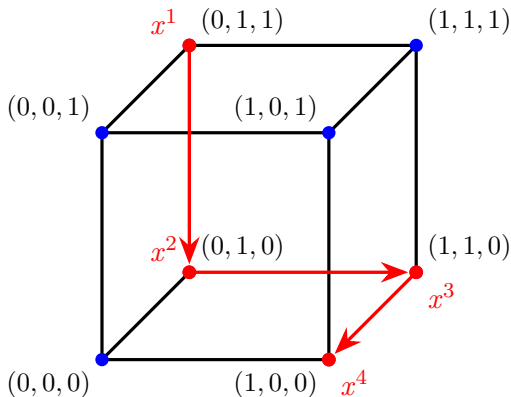
# Моделювання об'єкту дослідження

Еволюцію бінарної послідовності довжини  $N$  уявимо як випадкове блукання вершинами  $N$ -вимірного куба.



# Моделювання об'єкту дослідження

Ліниве блукання: з імовірністю  $p$  залишаємося на місці, з імовірністю  $\frac{1-p}{N}$  переходимо в сусідню вершину.



	$x^t$
$t = 1$	$(0, 1, 1)$
$t = 2$	$(0, 1, 0)$
$t = 3$	$(1, 1, 0)$
$t = 4$	$(1, 0, 0)$

# Моделювання об'єкту дослідження

$\{X^t\}_{t=\overline{1,T}}$  є ланцюгом Маркова зі станами в  $E = \{0, 1\}^N$  з початковим рівномірним розподілом  $\pi$  :

$$\pi_x = P(X^1 = x) = \frac{1}{2^N}$$

та матрицею перехідних імовірностей  $A$  :

$$A_{xx'} = P(X^{t+1} = x' \mid X^t = x) = \begin{cases} p, & d_H(x, x') = 0 \\ \frac{1-p}{N}, & d_H(x, x') = 1 \\ 0, & \text{інакше} \end{cases}$$



# Моделювання об'єкту дослідження

Спостерігаємо набір функціоналів

$$Y^t = (Y_1^t, \dots, Y_L^t) = \left( \sum_{i \in I_1} X_i^t, \dots, \sum_{i \in I_L} X_i^t \right),$$

де  $I_1, \dots, I_L$  є заданими підмножинами  $\{1, 2, \dots, N\}$ .

# Моделювання об'єкту дослідження

Спостерігаємо набір функціоналів

$$Y^t = (Y_1^t, \dots, Y_L^t) = \left( \sum_{i \in I_1} X_i^t, \dots, \sum_{i \in I_L} X_i^t \right),$$

де  $I_1, \dots, I_L$  є заданими підмножинами  $\{1, 2, \dots, N\}$ .

Наприклад,  $N = 12$ ,  $I_1 = (1, 2, 3)$ ,  $I_2 = (6, 7, 10, 11, 12)$

	$x^t$	$y^t$
$t = 1$	010011101101	(1, 4)
$t = 2$	011011101101	(2, 4)
$t = 3$	011011111101	(2, 4)
$t = 4$	011011111111	(2, 5)

# Моделювання об'єкту дослідження

## Твердження

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

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

# План доповіді

- 1 Сфери застосування бінарних послідовностей
- 2 Моделювання об'єкту дослідження
- 3 Побудова оцінок невідомих параметрів**
- 4 Результати чисельного експерименту

# Постановка задачі

- 1 Оцінити параметр  $p$  за набором спостережень та декодувати послідовність станів прихованого ланцюга;

# Постановка задачі

- 1 Оцінити параметр  $p$  за набором спостережень та декодувати послідовність станів прихованого ланцюга;

Метод максимальної правдоподібності

$$P(Y = y | p) = \sum_{x \in E^T} P(X = x, Y = y | p) \longrightarrow \max$$

Функція повної правдоподібності

$$L_{p,x,y} = P(X = x, Y = y | p)$$

Відтак

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

# Побудова оцінок невідомих параметрів

Ітераційний алгоритм Баума-Велша:

$$Q(p^{(n)}, p) = \sum_{x \in E^T} L_{p^{(n)}, x, y} \cdot \ln L_{p, x, y} \longrightarrow \max$$

Тож починаючи з деякого  $p^{(0)}$

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

# Побудова оцінок невідомих параметрів

Формула переоцінки параметра  $p$  :

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

де

$$\alpha_t(x) = P(Y^1 = y^1, \dots, Y^t = y^t, X^t = x | p^{(n)})$$

$$\beta_t(x) = P(Y^{t+1} = y^{t+1}, \dots, Y^T = y^T | X^t = x, p^{(n)})$$



# Побудова оцінок невідомих параметрів

Алгоритм Вітербі: пошук такої послідовності прихованих станів  $\hat{X}^1, \hat{X}^2, \dots, \hat{X}^T$ , яка найкращим чином описує наявні спостереження:

$$\hat{X} = \operatorname{argmax}_{x \in E^T} P(X = x \mid Y = y, \hat{p})$$

# Постановка задачі

- 2 Спостерігаємо значення  $Y_{I_*}^t = \sum_{i \in I_*} X_i^t$ , де  $I_*$  — деяка невідома підмножина множини індексів.

# Постановка задачі

- 2 Спостерігаємо значення  $Y_{I_*}^t = \sum_{i \in I_*} X_i^t$ , де  $I_*$  — деяка невідома підмножина множини індексів.

Наприклад,  $N = 12$ ,  $I_1 = (1, 2, 3)$ ,  $I_2 = (6, 7, 10, 11, 12)$ ,  $I_* = ?$

	$x^t$	$y^t$	$y_{I_*}^t$
$t = 1$	010011101101	(1, 4)	3
$t = 2$	011011101101	(2, 4)	3
$t = 3$	011011111101	(2, 4)	4
$t = 4$	011011111111	(2, 5)	4

# Постановка задачі

- 2 Спостерігаємо значення  $Y_{I_*}^t = \sum_{i \in I_*} X_i^t$ , де  $I_*$  — деяка невідома підмножина множини індексів.

Наприклад,  $N = 12$ ,  $I_1 = (1, 2, 3)$ ,  $I_2 = (6, 7, 10, 11, 12)$ ,  $I_* = ?$

	$x^t$	$y^t$	$y_{I_*}^t$
$t = 1$	010011101101	(1, 4)	3
$t = 2$	011011101101	(2, 4)	3
$t = 3$	011011111101	(2, 4)	4
$t = 4$	011011111111	(2, 5)	4

Яким чином можна відтворити елементи множини  $I_*$  за спостереженнями  $Y_{I_*}^1, \dots, Y_{I_*}^T$ ?

# Побудова оцінок невідомих параметрів

## Твердження

*Змістовною і незміщеною оцінкою потужності множини  $I_*$  є статистика*

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

# Побудова оцінок невідомих параметрів

Визначення компонент множини  $I_*$  :

$$\hat{I} = \operatorname{argmin}_{1 \leq k \leq C_N^{|\hat{I}_*|}} d \left( \hat{Y}_{I_k}, Y_{I_*} \right),$$

тут

$$\hat{Y}_{I_k} = \sum_{i \in I_k} \hat{X}_i^t$$

є сумою від декодованих елементів прихованого ланцюга.

# Побудова оцінок невідомих параметрів

Визначення компонент множини  $I_*$  :

$$\hat{I} = \operatorname{argmin}_{1 \leq k \leq C_N^{|\hat{I}_*|}} d \left( \hat{Y}_{I_k}, Y_{I_*} \right),$$

тут

$$\hat{Y}_{I_k} = \sum_{i \in I_k} \hat{X}_i^t$$

є сумою від декодованих елементів прихованого ланцюга.

Що обрати в ролі міри близькості  $d$ ?

# Побудова оцінок невідомих параметрів

## Середньоквадратична відстань

$$d_S \left( \widehat{Y}_{I_k}, Y_{I_*} \right) = \sum_{t=1}^T \left( \widehat{Y}_{I_k}^t - Y_{I_*}^t \right)^2$$

## Зважена відстань Жаккара

$$d_J \left( \widehat{Y}_{I_k}, Y_{I_*} \right) = 1 - \frac{\sum_{t=1}^T \min \left( \widehat{Y}_{I_k}^t, Y_{I_*}^t \right)}{\sum_{t=1}^T \max \left( \widehat{Y}_{I_k}^t, Y_{I_*}^t \right)}$$



# Постановка задачі

- 3 Спостереження на множинах  $I_1, \dots, I_L$  спотворюються ймовірностями  $q_1, \dots, q_L$  :

$$Y^t = (Y_k^t)_{k=\overline{1,L}} = \left( \sum_{i \in I_k} \tilde{X}_i^t \right)_{k=\overline{1,L}}$$

де для  $i \in I_k$

$$\tilde{X}_i^t = \begin{cases} 1 - X_i^t, & \text{з ймовірністю } q_k \\ X_i^t, & \text{з ймовірністю } 1 - q_k \end{cases}$$

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

# Побудова оцінок невідомих параметрів

Наприклад,  $N = 12$ ,  $I_1 = (1, 2, 3)$ ,  $I_2 = (6, 7, 10, 11, 12)$

	$x^t$	$\tilde{x}^t$	$y^t$	$q$
$t = 1$	010011101101	000011101101	$(0, 4)$	$(q_1, q_2)$
$t = 2$	011011101101	010011101101	$(1, 4)$	$(q_1, q_2)$
$t = 3$	011011111101	111011111111	$(3, 5)$	$(q_1, q_2)$
$t = 4$	011011111111	011011111100	$(2, 3)$	$(q_1, q_2)$

# Побудова оцінок невідомих параметрів

## Твердження

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

$$B_{xy}^q = P(Y^t = y \mid X^t = x) = \prod_{k=1}^L P(\xi_{01}^k(x) + \xi_{11}^k(x) = y_k),$$

$$\xi_{01}^k(x) \sim \text{Bin}\left(|I_k| - \sum_{i \in I_k} x_i, q_k\right), \quad \xi_{11}^k(x) \sim \text{Bin}\left(\sum_{i \in I_k} x_i, 1 - q_k\right), \quad k = \overline{1, L}$$

# Побудова оцінок невідомих параметрів

Починаючи з деякого наближення моделі  $(\pi, A^{(0)}, B^{q^{(0)}})$ , формула переоцінки параметра  $p$  :

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

Формула переоцінки компонент вектора  $q$  :

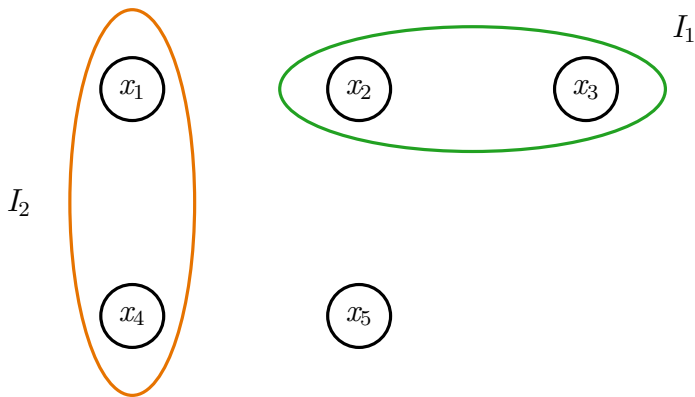
$$q_k^{(n+1)} = q_k^{(n)} \cdot \frac{\sum_{t=1}^T \sum_{x \in E} \beta_t(x) \sum_{x' \in E} \alpha_{t-1}(x') A_{x'x}^{(n)} \sum_{i \in I_k} P_{x,i}^{q^{(n)}}}{|I_k| \sum_{t=1}^T \sum_{x \in E} \alpha_t(x) \beta_t(x)}$$

# План доповіді

- 1 Сфери застосування бінарних послідовностей
- 2 Моделювання об'єкту дослідження
- 3 Побудова оцінок невідомих параметрів
- 4 **Результати чисельного експерименту**

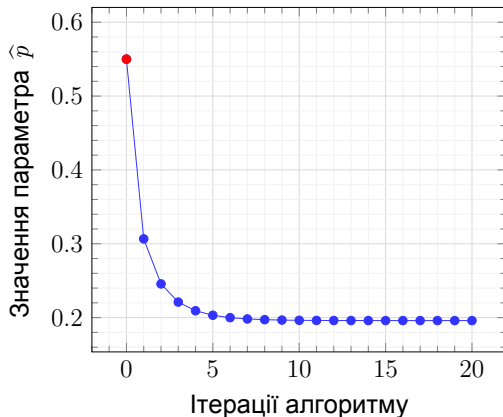
# Результати чисельного експерименту

Було згенеровано прихований ланцюг Маркова протягом  $T = 200$  моментів часу,  $N = 5$  та  $p = 0.2$ . Множина спостережуваних індексів:  $I = \{I_1, I_2\} = \{(2, 3), (1, 4)\}$ .



# Результати чисельного експерименту

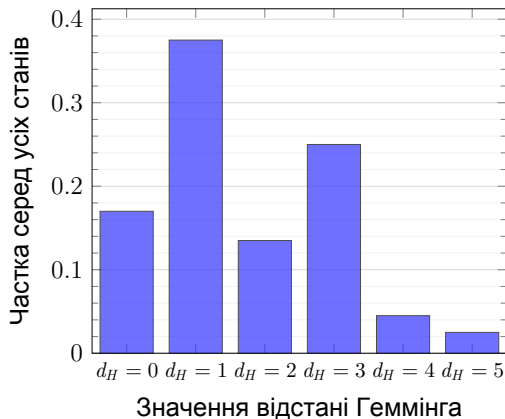
- 1 Оцінити параметр  $p$  за набором спостережень та декодувати послідовність станів прихованого ланцюга;



$p$	$\hat{p}$	$ p - \hat{p} $
0.2	0.1959	0.0041

# Результати чисельного експерименту

- 1 Оцінити параметр  $p$  за набором спостережень та декодувати послідовність станів прихованого ланцюга;



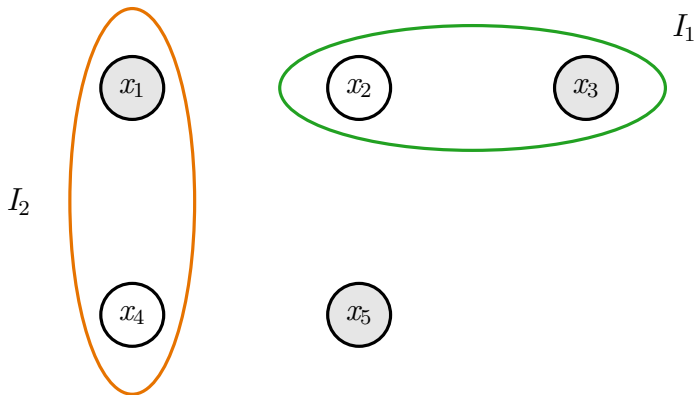
$$d_H(X^t, \widehat{X}^t) = \sum_{i=1}^N \mathbb{1}(X_i^t \neq \widehat{X}_i^t)$$



# Результати чисельного експерименту

2 Відтворити елементи «множини неявних індексів»  $I_*$ ;

Спостереження  $I = \{I_1, I_2\} = \{(2, 3), (1, 4)\}$ , неявні індекси покладемо  $I_* = (1, 3, 5)$  :



# Результати чисельного експерименту

- 2 Відтворити елементи «множини неявних індексів»  $I_*$ ;

Залежність значення оцінки від довжини ланцюга

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

# Результати чисельного експерименту

- 2 Відтворити елементи «множини неявних індексів»  $I_*$ ;

Залежність значення оцінки від довжини ланцюга

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

При малих  $N$  можна використати оцінку

$$|\widehat{I_*}|_{\max} = \max_{1 \leq t \leq T} Y_{I_*}^t$$

# Результати чисельного експерименту

- 2 Відтворити елементи «множини неявних індексів»  $I_*$ ;

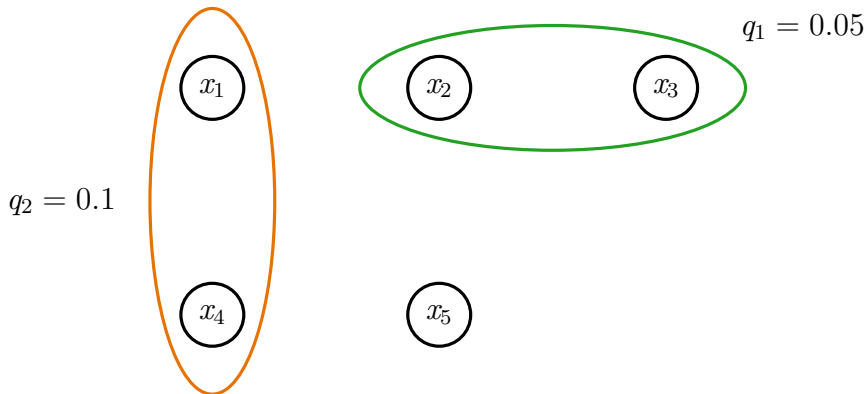
Отримані результати:

Істинна множина $I_*$	(1, 3, 5)
Оцінка $\hat{I}_S$ за середньоквадратичною відстанню	(1, 2, 5)
Оцінка $\hat{I}_J$ за зваженою відстанню Жаккара	(1, 2, 3)

# Результати чисельного експерименту

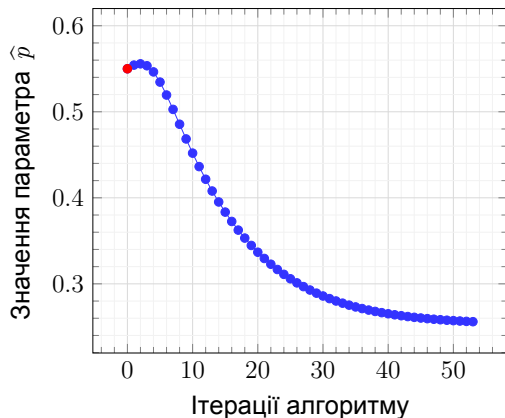
- 3 Оцінити невідомий параметр моделі  $p$  при ймовірностях спотворення  $q_1, q_2, \dots, q_L$ .

Для  $I_1, I_2$  було задано такі коефіцієнти спотворення:



# Результати чисельного експерименту

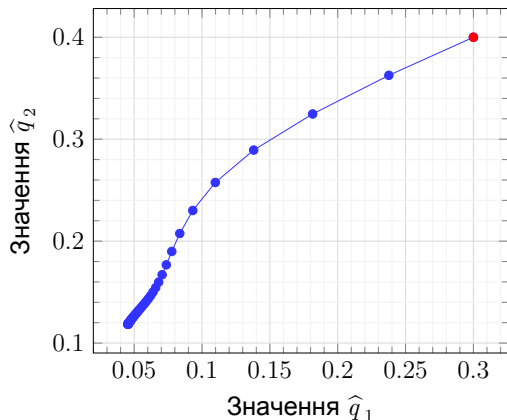
- 3 Оцінити невідомий параметр моделі  $p$  при ймовірностях спотворення  $q_1, q_2, \dots, q_L$ .



$p$	$\hat{p}$	$ p - \hat{p} $
0.2	0.2559	0.0559

# Результати чисельного експерименту

- 3 Оцінити невідомий параметр моделі  $p$  при ймовірностях спотворення  $q_1, q_2, \dots, q_L$ .



$q_1$	$\hat{q}_1$	$ q_1 - \hat{q}_1 $
0.05	0.0454	0.0046

$q_2$	$\hat{q}_2$	$ q_2 - \hat{q}_2 $
0.1	0.1184	0.0184

Невідомі параметри заданої 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 при збільшенні кількості спостережень.