Advanced ML & AGT

Class 11

Bandits v. Knapsacks

Bandits with Knapsacks

1. מוטיבציה ומודל
2. אלגוריתם מרכזי
3. חסם תחתון
4. אלגוריתם במו"מ UCB

מודיפיקציה: Dynamic Pricing

רוכשים לאורך B מוצרים בבים

בזמן t

בא לקוח (עם ערך v_t)

המוכר מציע מחיר p_t

אם $v_t \geq p_t$ יש קנייה

עבור $B=T$ (אין מגבלת כמות)

ראינו בעבר.

מה לעשות אם $B<T$?

עלולים להישאר ללא מוצרים

מספר מוצרים

d מוצרים

B_i עותקים של מוצר i

קונה: ערך $v_t = (v_{t,1}, \dots, v_{t,d})$

מוכר: מחירים $p_t = (p_{t,1}, \dots, p_{t,d})$

התוצאה:

נקנו המוצרים i עבורם

$$v_{t,i} \geq p_{t,i}$$

☆ אפשר גם מוכרים אחרים לקנייה

מודל Bandit with Knapsacks

הגדרה: קיים אילוצים שלובלים שצריך לקיים.

משאבים בעלי כמות מוגבלת

משקל i יש כמות B_i

רווחים רוצים להגיע למקסימום.

דינמיקה בכל זמן t

בוחרים פעולה a_t

מקבלים רווח r_t

צורכים משאבים

משאב i בכמות $0 \le c_{t,i}$

הבעיה פורמלית

פרמטרים: k פעולות

d משאבים

$B_1 \dots B_d$ תקציבים

T זמן

בזמן t:

בוחרים פעולה $a_t \in [k]$

רואים תוצאה $O_t = (r_t; c_{t,1}, \dots, c_{t,d}) \in [0,1]^{d+1}$

$r_t \in [0,1]$ רווח

$c_{t,i} \in [0,1]$ צריכת משאב i

עוצרים אם קיים משאב i עבורו

$\sum c_{t,i} > B_i$

מודל סטוכסטי:

לכל פעולה a יש התפלגות P_a

ודוגמים $o_t \sim P_a$

בזמן t דוגמים מטריצה

$M_t \in [0,1]^{K \times (d+1)}$ מהתפלגות P

השורה $M_t(a)$ המתאימה ל a

בזמן t

$$M_t(a) = (r_t(a); c_{t,1}(a), \dots, c_{t,d}(a))$$

כלומר $o_t = M_t(a_t)$

המטרה:

להביא למקס. את הרווח

$$\sum r_t$$

בזמן משאבה נופל

ירד ב-1 בכל פעם.

$\forall i \quad B_i = T$

'GO/GO MAB

דוגמאות

תמחור דינמי

T מספר לקוחות

B מספר אתקים

פעולות מחירים

לקוח יש ערך v_t
וקונה אם $v_t \geq p_t$

$$\vec{O}_t = \begin{cases} (P;1,1) & \text{מצב בקנה} \\ (0;0,1) & \text{לא בקנה} \end{cases}$$

תמחור דינמי לעלה דים

T זמן

B תקציב

פעולות תשלום (מחיר)

מטרה להגיע מקס. משימות

לעובד יש ערך מינימלי v_t
ומסכים אם $v_t \leq p_t$

$$\vec{O}_t = \begin{cases} (1;P,1) & \text{משימה בוצעה} \\ (0;0,1) & \text{אחרת} \end{cases}$$

Repeated auction

T זמן (לקוחות)

B עותק של מוצר

בעלה מחיר מינימום θ_t

p_t המחיר בו נמכר המוצר

(אם נמכר, לפחות θ_t)

$$O_t = \begin{cases} (p_t, 1, 1) & \text{מוצר נמכר במחיר } p_t \\ (0, 0, 1) & \text{אחרת} \end{cases}$$

Pay per click ad auction

T זמן (משתמשים)

B תקציב של הלקוח

בעלה להציג מודעה a

r_a תשלום = רווח

אם בוצע קליק

(בהסתברות q_a לא יבוצע!)

מטרה: לסיים את תקציב הלקוח

$$O_t = \begin{cases} (r_a, r_a, 1) & \text{אם היה קליק} \\ (0, 0, 1) & \text{אחרת} \end{cases}$$

Repeated Bidding on a Budget

בקרת המבט של קונה

T מספר המכרזים

B תקציב

פעולה b_t (bid למוצר)

מחיר $p_t \le b_t$

מטרה: לקנות מספר מקסימלי מוצרים בתקציב B

$$\vec{o}_t = \begin{cases} (1; p_t, 1) & \text{קיבל במחיר } p_t \\ (0; 0, 1) & \text{לא קיבל} \end{cases}$$

הסדרות על המודל

1) צריכת משאבים מקדימה (Explore-Exploit) עלולה להביא לחרטה.

2) הרווח לפעולה - לא המטרה

צריך לחשוב על סכום הרווחים

3) הפעולה הטובה ביותר, לא המצב

דוגמא: שתי פעולות, תקציב B כמו $T=2B$.

בכל כמו t, $b_t=1$

$a_1 \to (1;1,0,1)$

$a_2 \to (1;0,1,1)$

כל פעולה בודדת תשיג B

שימוש בשני הפעולות (לחילופין) ישיג $2B$!

Job scheduling

K מספר המכונות

לכל מכונה i נמשאבים:

CPU_i, $memory_i$, $bandwidth_i$

J מספר עבודות

כל עבודה j דורשת משאבים

רואים את הבעיות רק אחרי

התצמון על מכונה.

מטרה: להביא למקס. את מספר

העבודות שנגיש.

פעולות $a \in [K]$

תגובה $(a=i)$:

$$O_t = (0 \dots 0, \underbrace{CPU_t, mem_t, bw_t}_{\text{מכונה } i}, 0 \dots 0, \underset{\text{לזמן}}{1})$$

מס"מים כאשר אחת המכונות

"מלאה"

Contextual MAB:

ALG בוחר את הפעולה
הטובה ביותר

MAB w. Knapsacks

בעיה שמסיק לבחירת התפלגות
על פעולות

Benchmark

$$OPT \triangleq \sup_{ALG} E[REW(ALG|I)]$$

אלגוריתם

פרופיל
הבעיה

BwK

Linear Relaxation: $OPT_{LP} \geq OPT$

Linear Program

Lagrange func.: Game value $= OPT_{LP}$

Lagrange Game

Learning in Games: Avg Play $\approx$ Nash

Repeated Game

Large reward at stopping time

BwK

LagrangeBwK אלגוריתם

Linear Relaxation

נבנה תוכנית לינארית
התפלגות D מעל פעולות
נבנה $OPT_{LP} \geq \frac{OPT}{T}$

Lagrange Game

נבנה משחק סכום-אפס שקול ל LP
שחקן אחד בחר פעולה a
שחקן שני בחר משקל i
ערך המשחק $= OPT_{LP}$

BwK

Linear Relaxation: $OPT_{LP} \geq OPT$

Linear Program

Lagrange func.: Game value $= OPT_{LP}$

Lagrange Game

Learning in Games: Avg Play $\approx$ Nash

Repeated Game

Large reward at stopping time

BwK

Repeated Lagrange Game

משתנה תגמול $\mathcal{L}_t(a,i)$

בתוחלת במשתנה המקורי $E[\mathcal{L}_t] = \mathcal{L}$

נפתור ע"י אלג' regret

עבור משתנה סכום אפס

(גורם רטרוספקטיבי לregret)

Reward at stopping

הרווח בזמן העצירה

דומה לרווח המשתנה

(גורם שני לregret)

הקדמה

$r(a)$, $C_i(a)$ תמלית רווח וגבייה

$B = \min_{1 \le i \le d} B_i$ נמלוך כל משלמה i ב- B_i/B

התקציב של כל משלמה B

המשמן "נגרק" הקצבה B/T

null arm רווח אפס

גבייה אפס, מלבד המשמן

Linear Relaxation

D התפלגות על פעולות

$$r(D) = \sum_a D(a)\, r(a) \qquad c_i(D) = \sum_a D(a)\, c_i(a)$$

נשתמש ב-D בכל זמן
רוחב וצריכה לפי התוחלת
תחרותית לאלגוריתם

$$\max \; r(D)$$
$$\text{s.t.} \quad D \in \Delta([k])$$
$$T\, c_i(D) \le B \qquad \forall i \in [d]$$

ערך ה LP הוא $OPT_{LP} \ge 1$

טענה: $OPT_{LP} \ge OPT$

הוכחה: ALG אלגוריתם לBwK

T עם null action. נסמן $\tau < T$ זמן עצירה של ALG,

D התפלגות אימפירית של ALG

$$D(a) = E\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{a_t = a\}\right] = \frac{1}{T}\sum_{t=1}^{T} \Pr[a_t = a]$$

$$E[REW(ALG)] = r(D)\cdot T$$ מתקיים

$$E[r_t] = \sum_a \Pr[a_t = a]\; E[r_t \mid a_t = a]$$

$$E[REW] = E\left[\sum_{t=1}^{T} r_t\right] = \sum_a r(a) \underbrace{\sum_t \Pr[a_t = a]}_{T\cdot D(a)} = T\cdot r(D)$$

המשך הוכחה:

באופן דומה $\sum_{t=1}^{T} E[c_{t,i}] = T \cdot c_i(D)$

$$E[c_{t,i}] = \sum_a Pr[a_t = a]\, E[c_{t,i} \mid a]$$

$$\sum_{t=1}^{T} E[c_{t,i}] = \sum_a \underbrace{\sum_{t=1}^{T} Pr[a_t = a]}_{T\, D(a)}\, c_i(a)$$

$$= T \cdot c_i(D)$$

כיוון שהאלגוריתם עובד רק בזמן T

$$T \cdot c_i(D) \le B$$

לכן D פתרון feasible של ה- LP

מכאן

$$E[REW(ALG)] = r(D) \le OPT_{LP}$$

$\star$

פונקציית Lagrange

פונקציית Lagrange של בעיית LP

$$\mathcal{L}(D,\lambda) = r(D) + \sum_{i\in[d]} \lambda_i \left(1 - \frac{I}{B} c_i(D)\right)$$

לפי זיר משתמש סכום לפס:

$$\mathcal{L}(a,i) = r(a) + 1 - \frac{I}{B} c_i(a)$$

שמקבל max בוחר פעולה a

שמקבל min בוחר משאב i

$$\min_{\lambda} \max_{D} \mathcal{L}(D,\lambda) = \max_{D} \min_{\lambda} \mathcal{L}(D,\lambda) = OPT_{LP}$$

למה: עבור שיווי משקל Nash (D^*,λ^*) מתקיים:

(א) $0 \le 1 - \frac{I}{B} C(D^*)$ ושיווי לא $0 < \lambda_i^*$

(ב) D^* פתרון אופטימלי של בעיית LP

(ג) $OPT_{LP} = \mathcal{L}(D^*,\lambda^*)$

הוכחה: עבור Nash מתקיים

$$\mathcal{L}(D^*,\lambda) \ge \mathcal{L}(D^*,\lambda^*) \ge \mathcal{L}(D,\lambda^*)$$

לפי זיר $Y_i = 1 - \frac{I}{B} c_i(D^*)$

נוכיח את (א) ע"י מקרים:

(i) עבור $\lambda_i^* = 1$ מתקיים $Y_i \ge 0$

(ii) $Y_i \ge 0$ (כאשר $\lambda_i^* < 1$)

(iii) $Y_i > 0$ ו- $0 < \lambda_i^*$ בלתי אפשרי.

הוכחת הלמה חלק ב:

(i) אם $\lambda_i^* = 1$ אזי $Y_i = 1 - \frac{T}{B} C_i(D^*) \geq 0$

אם זה בדיוק האמצע, אזי $\frac{B}{T} = C_i(D^*)$ ו- $Y_i = 0$

נקבע פעולה a, ונגדיר התפלגות D

$D(null) = D^*(null) + D^*(a)$ $\quad D(a') = D^*(a')$ $\quad D(a) = 0$
$a' \neq a$

$$0 \leq \mathcal{L}(D^*, \lambda^*) - \mathcal{L}(D, \lambda^*)$$
$$= [r(D^*) - r(D)] + \frac{T}{B}[C_i(D^*) - C_i(D)]$$
$$= D^*(a)\left[r(a) - \frac{T}{B} C_i(a)\right]$$
$$\leq D^*(a)\left[1 - \frac{T}{B} C_i(a)\right]$$

נסכום על הפעולות ונקבל $Y_i \geq 0$.

(ii) $Y_i \geq 0$ (מספיק להראות עבור $\lambda_i^* < 1$)

בשלילה, נניח משהו i עם $\lambda_i^* < 1$ ו- $Y_i < 0$ מינימלי

נגדיר פילוג λ: $\lambda_i = 1$, $\lambda_{i'} = 0$ $(i' \neq i)$

נקבל $\mathcal{L}(D^*, \lambda) < \mathcal{L}(D^*, \lambda^*)$, סתירה!

(iii) אם $Y_i > 0$ אזי $\lambda_i^* = 0$

נניח $\lambda_i^* > 0$ ו- $Y_i > 0$

$$r(D^*) < \mathcal{L}(D^*, \lambda^*)$$

נבחר λ כך: $\lambda(null) = 1$, $\lambda(a) = 0$ $(a \neq null)$

$$\mathcal{L}(D^*, \lambda^*) > r(D^*) = \mathcal{L}(D^*, \lambda)$$

סתירה!

הוכחת הלמה

תלקים ב, ג:

נתבונן ב

D^* פתרון feasible של ה-LP

$$r(D^*) = \mathcal{L}(D^*, \lambda^*)$$

לכל D מתקיים

$$r(D^*) = \mathcal{L}(D^*, \lambda^*) \geq \mathcal{L}(D, \lambda^*) \geq r(D)$$

נתבונן D^* פתרון אופטימלי של LP

$$OPT_{LP} = r(D^*) = \mathcal{L}(D^*, \lambda^*)$$

Repeated Lagrange Game

בצעד t יש מכרז M_t מאבדה $\mathcal{L}_t$

$$\mathcal{L}_t(a,i) = v_t(a) + 1 - \frac{T}{B} c_{t,i}(a)$$

מתקיים $E[\mathcal{L}_t(a,i)] = \mathcal{L}(a,i)$

משחק

ALG-Max בוחר פעולה a_t

ALG-Min בוחר מטלה i_t

התגמולה $\mathcal{L}_t(a_t, i_t)$

תשובה:

יש מספיק מנגנון למצוא עיצוב המשחק!

אלגוריתם LagrangeBwK

קלט: T, B, K, d

ALG-Max בוחר פעולה (מן פל bandit)

ALG-Min בוחר מטלה (מן פל full-info)

בצעד t:

בעקבות: ALG-Max בוחר $a_t \in [K]$

ALG-Min בוחר $i_t \in [d]$

פעולה a_t מתבצעת

רושם $O_t = (r_t(a_t); c_{t,1}(a_t), \dots)$

ALG-Max רושם $\mathcal{L}_t(a_t, i_t)$

ALG-Min רושם $\mathcal{L}_t(a_t, i)$ $\forall i \in [d]$

ניתוח:

ALG-max ל' regret של $R_1(T)$

פעולה ממוצעת $\bar{a}_\tau = \frac{1}{\tau}\sum_{t\in[\tau]} a_t$

ALG-min ל' regret של $R_2(T)$

משקל ממוצע $\bar{\mu}_\tau = \frac{1}{\tau}\sum_{t\in[\tau]} i_t$

מתמשתקים סכום טופס

בזמן $\tau\in[T]$

הממוצע הלוגריתמי $(\bar{a}_t, \bar{\mu}_t)$

הוא δ_τ-Nash כלשהי

$$\tau\,\delta_\tau = R_1(\tau) + R_2(\tau) + err_\tau$$

$$err_\tau = \left|\sum_{t\in[\tau]} \mathcal{L}_t(i_t, j_t) - \mathcal{L}(i_t, j_t)\right|$$

כלומר

$$\forall i\in[d] \quad \mathcal{L}(\bar{a}_\tau, i) \geq OPT_{LP} - \delta_\tau$$

הוכחה: מתקיים

$$\sum_{t\in[\tau]} c_{t,i} > B$$

לכן

$$\tau \mathcal{L}(\bar{a}_\tau, i) = \tau\left(r(\bar{a}_\tau) + 1 - \frac{T}{B} c_i(\bar{a}_\tau)\right)$$

$$\le \sum_{t\in[\tau]} r_t + \tau - \frac{T}{B}\underbrace{\sum_{t\in[\tau]} c_{i,t}}_{B} + err^*_{\tau,i}$$

$$\le REW - (T-\tau) + err^*_{\tau,i}$$

$$REW(ALG) \ge \tau \mathcal{L}(\bar{a}_\tau, i) + (T-\tau)\underbrace{OPT_{LP}}_{\le 1} - err^*_{\tau,i}$$

Reward at Stopping

בזמן τ יש משלב i שנגמר

למה:

$$REW(ALG) \ge \tau \mathcal{L}(\bar{a}_\tau, i) + (T-\tau) OPT_{LP} - err^*_{\tau,i}$$

$$err^*_{\tau,i} = \left|\tau \cdot r(\bar{a}_\tau) - \sum_{t\in[\tau]} r_t\right| + \frac{T}{B}\left|\tau \cdot c_i(\bar{a}_\tau) - \sum_{t\in[\tau]} c_{i,t}\right|$$

לכן אם אלגוריתם ALG

לפי אי שוויון הוֹפדינג:

$$err_{\tau,i} \le \sqrt{TK \log \frac{dT}{\delta}} \qquad \forall \tau \in [T],\ i \in [d]$$

3 כיוון אלגו' regret יש הסתברות גבוהה

$$R_1(\tau) \le z\sqrt{TK \log(T/\delta)}$$

$$R_2(\tau) \le z\sqrt{T \log(dT/\delta)}$$

משפט

$$\underbrace{OPT - REW}_{\text{regret}} \le O\left(\frac{T}{B} \cdot \sqrt{TK \log(dT/\delta)}\right)$$

חסם הדוק עבור $B = \Omega(T)$

חסם לא טריוויאלי עבור $B = \Omega(\sqrt{T})$

סיכום

$$\mathcal{L}(\bar{a}_\tau, i) \ge OPT_{LP} - \delta_\tau \qquad \forall i$$

$$\tau \delta_\tau = R_1(\tau) + R_2(\tau) + err_\tau$$

$$Rew \ge \tau \mathcal{L}(\bar{a}_\tau, i) + (T-\tau) OPT_{LP} - err_\tau$$

$$Rew \ge T \cdot OPT_{LP} - R_1(\tau) - R_2(\tau) - err_\tau$$

נתון (בהסתברות גבוהה):

$$err_\tau \le z\sqrt{TK \log\left(\frac{dT}{\delta}\right)}$$

$$z = \max |\mathcal{L}(a,i)| \le \frac{T}{B}$$

$$\mathcal{L}(a,i) \in [1 - \tfrac{T}{B}, 2]$$

חסם תחתון (בקיצור)

משפט:

$$\text{Regret} = \Omega\left(\min\left\{OPT, \sqrt{k\,OPT} + OPT\sqrt{\frac{k}{B}}\right\}\right)$$

חסם תחתון לכל $\min\{OPT, \sqrt{k\,OPT}\}$

נבנה שירות מ- MAB

חסם תחתון $\min\{OPT, OPT\sqrt{k/B}\}$

הרעיון: דומה ל MAB

הרווח תמיד 1.

הבריכה $\{0,B\}$

פעולה אופטי $\Pr[c(a^*)=1] = \frac{1-\varepsilon}{2}$

פעולה אחרת $\Pr[c(a)=1] = \frac{1}{2}$

הזמן $T = 4B$ (לא אינפי)

תוחלת OPT $2B/(1-\varepsilon)$

<u>הערה</u>: ההבדל מ MAB

הזמן הוא מינ

הרווחים דטרמיניסטים

שימוש בחסם MAB

נקבע $T = 2B - \sqrt{B}\log 1/\varepsilon$, $T \le c\frac{k}{\varepsilon^2}$

בהסתברות גבוהה לא נעבור לפני T

עבור פרופיל רנדומלי

בכל צעד t,

בהסתברות $1/2$ נשחק פעולה לא אופט.

נובע מהחסם התחתון ב MAB!

בצעד T ההפרש בתכיכה

של OPT לעומת ONLINE

$\Theta(\varepsilon \cdot T)$

עבור $\varepsilon = \Theta(\sqrt{k/B})$ נקבל $\Theta(\sqrt{\frac{k}{B}}\underbrace{OPT}_{\Theta(T)})$

ניתן שאתה צעד T אלגו ONLINE

משתמש פעולה אופטימלית

אם ל-OPT נשארו X_1 משאבים

ול-ONLINE נשארו X_2 משאבים

אזי תוחלת ה regret

$$\frac{X_1}{(1-\varepsilon)B} - \frac{X_2}{(1-\varepsilon)B} = \Theta(X_1 - X_2) = \Theta(\varepsilon T)$$

לכן $\text{regret} = \Theta\left(\sqrt{\frac{k}{B}}\, OPT\right) = \Theta(\varepsilon T)$

אלגוריתם UCB (המקיצור)

נראה גבול עליון ל UCB

$$\forall i,t \quad UCB_t(a_i) \geq \mu_i \geq LCB_t(a_i)$$

$$\left| \sum_{t=1}^{T} UCB_t(a_t) - r_t \right|$$

$$\leq \sum_{a} \left| \sum_{t: a_t = a} UCB_t(a_t) - LCB_t(a_t) \right|$$

$$\leq \sqrt{KT}$$

$$a_t = \arg\max_a UCB_t(a)$$

$$UCB_t(a_t) \geq UCB_t(a^*) \geq \mu^*$$

$$\text{Regret} = \sum_{t=1}^{T} \mu^* - r_t$$

$$\leq \sum_{t=1}^{T} UCB_t(a_t) - r_t$$

$$\leq \sqrt{KT}$$

תזכורת $LP(r, G, \varepsilon)$

$$\max_D \; r(D)$$

$$\text{s.t.} \quad G D \preceq B(1-\varepsilon)\vec{1}$$

$$D \in \Delta([K])$$

$$G = \begin{bmatrix} - & c_1 & - \\ - & c_d & - \end{bmatrix}$$

לשים ב D_t:

$$D_t \leftarrow LP(UCB_t(r), LCB_t(G), \varepsilon)$$

מ.ש.ל: $\vec{a} \succeq \vec{b} \iff \forall i \; a(i) \geq b(i)$

$\gamma = \log \frac{KTd}{\delta}$ $\quad \varepsilon = \sqrt{\frac{\gamma K}{B}} + \frac{\gamma K}{B} \log T$ עבור: נשים

$$\text{Regret} = \tilde{O}\left(OPT\sqrt{\frac{K}{B}} + \sqrt{K\, OPT} + K\right)$$

נשים את התסם התחתון!

הערה: בהסתברות גבוהה

$UCB_t(r) \succeq r \qquad LCB(G) \preceq C$

$$\left|\sum_{t=1}^{T} UCB_t(r) \cdot D_t - r_t\right| = O\left(\sqrt{\gamma K \sum r_t}\right)$$

$$\left|\sum_{t=1}^{T} LCB_t(G) D_t - c_t\right| \preceq \varepsilon \cdot B \vec{1}$$

(בכיף לתוכית ...)

$$REW(ALG) = \sum_{t=1}^{T} r_t$$

$$= \left[\sum_t UCB_t(r) D_t\right] - \left[\sum_t UCB_t(r) D_t - r_t\right]$$

$$\geq (1-\varepsilon) OPT_{LP} - \sqrt{\gamma K \cdot REW} - \gamma m$$

$$\text{Regret} = O\left(\varepsilon \cdot OPT + \sqrt{\gamma K \, OPT} + \gamma K\right)$$

$$\sqrt{\frac{K\gamma}{B}} \, OPT$$

$$\frac{1}{T}\sum_{t=1}^{T} UCB_t(r) \cdot D_t = \frac{1}{T}\sum_{t=1}^{T} LP(UCB_t(r), LCB_t(G), \varepsilon)$$

$$\geq LP(r, G, \varepsilon)$$

$$\geq (1-\varepsilon) OPT_{LP}$$

$$\frac{1}{T}\sum_{t=1}^{T} LCB_t(G) D_t \leq (1-\varepsilon) B \vec{1}$$

לכן

$$\frac{1}{T}\sum_{t=1}^{T} C_t \leq B\vec{1}$$

בהסתברות גדול יותר לפני זמן T

סיכום הרצאה

מודל BwK

אלגו מתבסס Lagrange

בקיצור:

חסם תחתון

אלגו דמוי UCB

מקורות

רוב ההרצאה

Slivkins, 10 פרק

חסם תחתון

Badanidiyuru, Kleinberg, Slivkins, 2018

(Bandits with knapsacks)

אלגוריתם דמוי UCB

Agrawal, Devanur, 2014