Δευτέρα 4/12, Θεωρία Παιχνιών (Βασίλης Σταμάτης)

Αναθεώρηση - Μοντέλα επαναλαμβανόμενης μάθησης (online learning) συνεχούς χρόνου.

- Κάθε χρ. στιγμή → διάνυσμα πληρωμών $v_t \in \mathbb{R}^A$
  → Επιλέγω μεικτή στρατηγική $x_t \in \Delta(A)$
  → Λαμβάνω $u_t(x_t) = \langle v_t, x_t \rangle$

- Μεταμέλεια (Regret): $Reg(T) = \max_{p \in \Delta(A)} \int_0^T [u_t(p) - u_t(x_t)]\,dt$.

- Αν ο παίκτης ακολουθεί (εω), τότε: $\boxed{Reg(T) \le \log A}$.

Online Learning σε διακριτό χρόνο.

Sequence of events Για κάθε χρ. στιγμή $t = 1, 2, \dots$

- Επιλογή μεικτής στρατηγικής $x_t \in \Delta(A)$.
- Επιλογή καθαρής στρατηγικής $a_t \sim x_t$. • Διάνυσμα πληρωμής: $v_t \in [0,1]^A$.
- Πληρωμή $u_t(a_t) = v_{t,a_t}$

Repeat.

Μεταμέλεια/Regret: $\sum_{t=1}^{T} [u_t(a) - u_t(a_t)] = RReg_a(T)$ → Realized Regret. [Στοχαστικά].

$\mathbb{E}_{a_t \sim p \in \Delta(A)} [u_t(a_t)] - \mathbb{E}_{a_t \sim x_t \in \Delta(A)} (u_t(a_t))$ ← Μέση Διαφορά.

$= \langle v_t, p \rangle - \langle v_t, x_t \rangle$

$Reg_p(T) = \sum_{t=1}^{T} \langle v_t, p - x_t \rangle = \sum_{t=1}^{T} u_t(p) - u_t(x_t)$

Έχουμε φτάσει στον ορισμό: $Reg_p(T) = \sum_{t=1}^{T} \mathbb{E}\left[\sum_{t=1}^{T} [u_t(p) - u_t(a_t)]\right]$!

Συνολική Μεταμέλεια (Worst-Case Regret)

$\boxed{Reg(T) = \max_{p \in \Delta(A)} \sum_{t=1}^{T} \langle v_t, p - x_t \rangle}$. Ποιο το πρόβλημα ?

Αν ξεκινήσουμε με τον ορισμό του realized regret:

$RRegret(T) = \sum_{t=1}^{T} [u_t(a) - u_t(a_t)]$

Θα έχουμε ως προς μία μεικτή $p \in \Delta(A)$

$RReg_p(T) = \sum_{t=1}^{T} [u_t(p) - u_t(a_t)]$

Οπότε, η χειρότερη περίπτωση θα είναι: $\max_{p\in\Delta(A)} RReg_p(T) = \max_{p\in\Delta(A)} \sum_{t=1}^{T}[u_t(p) - u_t(a_t)]$

Μέσο Regret $= \mathbb{E}_{a_t\sim x_t}\left[\max_{p\in\Delta(A)} \sum_{t=1}^{T}[u_t(p) - u_t(a_t)]\right]$.

• Υπάρχει διαφορά αν κοιτάξουμε:

① Την μέση τιμή ενός μεγίστου (mean regret)

② Το μέγιστο μίας μέσης τιμής (pseudo-regret)

Εμείς θα μελετήσουμε αποκλειστικά το 2ο δηλ.:

$$Reg(T) = \max_{p\in\Delta(A)} \sum_{t=1}^{T} \langle v_t, p - x_t\rangle$$

ή στην στοχαστική περίπτωση:

$$\overline{Reg}(T) = \max_{p\in\Delta(A)} \mathbb{E}\left[\sum_{t=1}^{T} \langle v_t, p - x_t\rangle\right]$$

Πληροφόρηση του Παίκτη: Αρκετές διαφορετικές περιπτώσεις:

① Πλήρης πληροφόρηση (full information).

Παρατηρούμενη ποσότητα: $v_t \in [0,1]^t$.

② Ενθόρυβες παρατηρήσεις πληρωμών:

Παρατηρούμενη ποσότητα: $\hat{v}_t = v_t + z_t$ ← δείγμα τυχαίου θορύβου.

③ "Bandit" / Παρατήρηση πραγματικής πληρωμής

Παρατηρούμενη ποσότητα: $u_t(a_t) = v_{t,a_t} \in [0,1]$

Παράδειγμα: R-P-S

Παίκτης A (εξτρεμιστής) παίξει στρατηγική $x_t = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$.

Παίκτης B (περιβάλλον): $y_t = (1/2, 1/2, 0)$.

$$v_t = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} \cdot y_t = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1/2 \\ 1/2 \\ 0 \end{pmatrix} = \begin{pmatrix} -1/2 \\ 1/2 \\ 0 \end{pmatrix}.$$

Περίπτωση 1: R (-½) P (½) S (0).

Περίπτωση 2: $z_R = 0{,}02$ $z_P = 0{,}03$ $z_S = 0{,}5$

R (-0,48) P (0,53) S (-0,5)

Περίπτωση 3: Ο παίκτης A παίζει

$a_t = (0,1,0) \overset{x_t}{\rightsquigarrow}$ Paper

Άρα παρατηρεί X ½ X.

Γενική μορφή πληροφορίας πρώτης τάξης: Την χρονική στιγμή t, ο παίκτης παρατηρεί το σήμα $\hat{v}_t = v_t + U_t + b_t$ όπου: $\mathbb{E}[U_t | \mathcal{F}_t] = 0$

U_t: τυχαίο σφάλμα, b_t: συστηματικό σφάλμα (bias)

$\mathcal{F}_t$: παρελθόν/ιστορία της διαδικασίας μέχρι τη στιγμή t.

$\rightarrow b_t = \mathbb{E}[\hat{v}_t | \mathcal{F}_t] - v_t$

Υποθέσεις: ① Συστηματικό σφάλμα: $\|b_t\| \le B_t$.

② Διασπορά τυχαίου σφάλματος: $\mathbb{E}[\|v_t\|^2 | \mathcal{F}_t] \le \sigma_t^2$.

③ Δεύτερη ροπή σήματος: $\mathbb{E}[\|\hat{v}_t\| | \mathcal{F}_t] \le M_t^2$.

Πλάνο: ① Θα πράξουμε ένα φράγμα μεταμέλειας (Regret Bound) για τον ΕΩ ως προς B_t, σ_t, M_t.

② Θα το εφαρμόσαμε σε κάθε μοντέλο πληροφόρησης.

Αλγόριθμος Hedge / Exponential Weights

$y_t = v_t$ (Συνεχής χρόνος)

$y_{t+1} = y_t + \gamma v_t$ (Διακριτός χρόνος)

$$x_t = \Lambda(y_t) = \frac{(\exp(y_{1,t}), \dots, \exp(y_{n,t}))}{\sum_{\beta \in A} \exp(y_{\beta,t})}$$

Τι Regret εξασφαλίζει ο Hedge?

Η λογική θα είναι ακριβώς όπως πριν κ' θα βασιστεί ε των συνάρτηση δυναμικού.

$\Phi(y) = \log \sum_{\alpha \in A} \exp(y_\alpha)$ (softmax / log-sum-exp)

Βασική ιδιότητα: $\nabla \Phi(y) = \Lambda(y)$.

Βήμα 1: $y_{t+1} = y_t + \gamma \hat{v}_t$. Θα μεζητήσουμε την αλλαγή $\Phi(y_{t+1})$ σε σχέση με το $\Phi(y_t)$.

Λήμμα: Έστω $y^+ = y + w$. Τότε, $\Phi(y^+) \le \Phi(y) + \langle \Lambda(y), w \rangle + \frac{1}{2} \|w\|_\infty^2$, $\|w\|_\infty = \max_{\alpha \in A} |w_\alpha|$

Απόδειξη: Από Taylor (με υπόλοιπο Lagrange): $\Phi(y^+) = \Phi(y) + \langle \nabla \Phi(y), y^+ - y \rangle + \frac{1}{2}(y^+ - y)^T \nabla^2 \Phi(y')(y^+ - y)$.

για κάποιο $y' \in [y, y^+]$.

① $= \langle \Lambda(y), y^+ - y \rangle = \langle \Lambda(y), w \rangle$

Για τον ② θα πρέπει να υπολογίσουμε τον Hessian με στοιχεία: $\frac{\partial^2 \Phi}{\partial y_\alpha \partial y_\beta} = \frac{\partial}{\partial y_\beta} \frac{\exp(y_\alpha)}{\sum_\gamma \exp(y_\gamma)}$

$= \dots = \chi_\alpha (\delta_{\alpha\beta} - \chi_\beta)$

↳ Kronecker $\delta_{\alpha\beta} = \begin{cases} 0, & \alpha \neq \beta \\ 1, & \alpha = \beta \end{cases}$. αναγέρεται στο γ.

(II) $\Rightarrow \sum_{\alpha,\beta} \chi_\alpha^{(2)} (\delta_{\alpha\beta} - \chi_\beta^{(1)}) w_\alpha w_\beta$

$= \sum_\alpha \chi_\alpha' w_\alpha^2 - \left(\sum_\alpha \chi_\alpha' w_\alpha\right)^2 \leq \sum_\alpha \chi_\alpha' w_\alpha^2 \leq \left(\sum_\alpha \chi_\alpha'\right) \max_\beta w_\beta^2 \leq \|w_\beta\|_\infty^2$. ($\sum_\alpha \chi_\alpha' = 1$)

Βήμα 2: Άρα, $\Phi(y+1) \leq \Phi(y_t) + \langle \nabla(y_t), y_{t+1} - y_t \rangle + \frac{1}{2}\|y_{t+1} - y_t\|_\infty^2$

$= \Phi(y_t) + \gamma \langle x_t, \hat{v}_t \rangle + \frac{\gamma^2}{2}\|\hat{v}_t\|_\infty^2 = \Phi(y_t) + \gamma \langle x_t, \hat{v}_t \rangle + \gamma \langle x_t, v_t \rangle +$

$+ \gamma \langle x_t, b_t \rangle + \frac{\gamma^2}{2}\|\hat{v}_t\|_\infty^2$

$\gamma \langle v_t, p - x_t \rangle = \Phi(y_t) - \Phi(y_{t+1}) + \gamma \langle b_t, x_t \rangle + \gamma \langle v_t, x_t \rangle + \frac{\gamma^2}{2}\|\hat{v}_t\|_\infty^2$.

$+ \gamma \langle v_t, p \rangle$.

Cauchy-Schwartz

Βήμα 3: Τηλεσκοπούμε: $\gamma \sum_{t=1}^T \langle v_t, p - x_t \rangle \leq \Phi(y_t) - \Phi(y_{t+1}) + 2\gamma \sum_{t=1}^T B_t$

$+ \gamma \sum_{t=1}^T \langle v_t, x_t \rangle + \frac{\gamma^2}{2} \sum_{t=1}^T \|\hat{v}_t\|_\infty^2$

$+ \gamma \sum_{t=1}^T \langle v_t, p \rangle$

$\mathbb{E}\left[\sum_{t=1}^T \langle v_t, p - x_t \rangle\right] \leq \frac{\Phi(y_t)}{\gamma} + \mathbb{E}\left[\langle y_{t+1}, p \rangle - \Phi(y_{t+1})\right] + \sum_{t=1}^T B_t$

$\overline{\text{Reg}_p(T)}$

$+ 0 + \frac{\gamma}{2} \sum_{t=1}^T M_t^2$