ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΠΑΙΓΝΙΩΝ ΚΑΙ ΛΗΨΗΣ ΑΠΟΦΑΣΕΩΝ

ΛΗΣΤΕΣ, ΚΟΥΛΟΧΕΡΗΔΕΣ, ΚΑΙ ΘΕΩΡΙΑ ΜΑΘΗΣΗΣ

Παναγιώτης Μερτικόπουλος

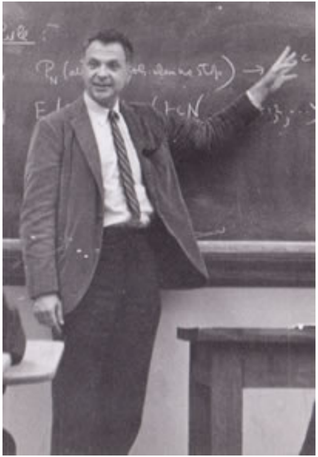
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

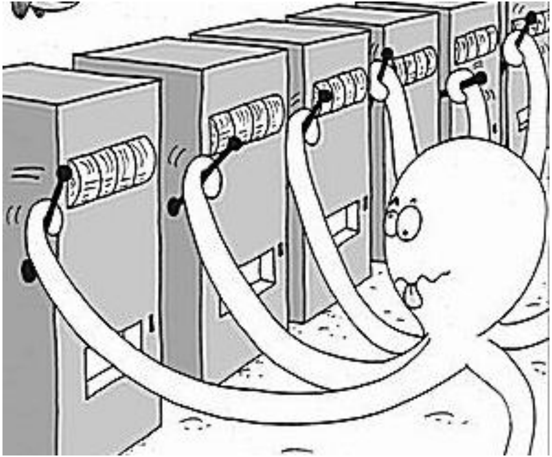
Τμήμα Μαθηματικών

Χειμερινό Εξάμηνο, 2023–2024

Multi-armed bandits

Robbins' multi-armed bandit problem: **how to play in a (rigged) casino?**





Outline

1. Online learning in continuous time
2. Online learning in discrete time
3. Learning with oracle feedback
4. Learning with bandit feedback

Oracle feedback

The oracle model

A ***stochastic first-order oracle (SFO)*** model of v_t is a random vector $\hat{v}_t$ of the form

$$\hat{v}_t = v_t + U_t + b_t \tag{SFO}$$

where U_t is **zero-mean** and $b_t = \mathbb{E}[\hat{v}_t \mid \mathcal{F}_t] - v(x_t)$ is the **bias** of $\hat{v}_t$

Oracle feedback

The oracle model

A ***stochastic first-order oracle (SFO)*** model of v_t is a random vector $\hat{v}_t$ of the form

$$\hat{v}_t = v_t + U_t + b_t \tag{SFO}$$

where U_t is **zero-mean** and $b_t = \mathbb{E}[\hat{v}_t \mid \mathcal{F}_t] - v(x_t)$ is the **bias** of $\hat{v}_t$

Assumptions

- ***Bias:*** $\|b_t\|_\infty \leq B_t$
- ***Variance:*** $\mathbb{E}[\|U_t\|_\infty^2 \mid \mathcal{F}_t] \leq \sigma_t^2$
- ***Second moment:*** $\mathbb{E}[\|\hat{v}_t\|_\infty^2 \mid \mathcal{F}_t] \leq M_t^2$

Oracle feedback

The oracle model

A ***stochastic first-order oracle (SFO)*** model of v_t is a random vector $\hat{v}_t$ of the form

$$\hat{v}_t = v_t + U_t + b_t \tag{SFO}$$

where U_t is **zero-mean** and $b_t = \mathbb{E}[\hat{v}_t \mid \mathcal{F}_t] - v(x_t)$ is the **bias** of $\hat{v}_t$

Algorithm HEDGE-O # ExpWeight with SFO feedback

Require: set of actions $\mathcal{A}$; sequence of payoff vectors $v_t \in \mathbb{R}^{\mathcal{A}}$, $t = 1, 2, \ldots$

Initialize: $y_1 \in \mathbb{R}^{\mathcal{A}}$

$$\Lambda(y) = \frac{(\exp(y_1), \ldots, \exp(y_A))}{\sum_\alpha \exp(y_\alpha)}$$

for all $t = 1, 2, \ldots$ **do**

  set $x_t \leftarrow \Lambda(y_t)$ # mixed strategy

  play $\alpha_t \sim x_t$ and **receive** $v_{\alpha_t, t}$ # choose action / get payoff

  observe $\hat{v}_t \leftarrow v_t$ # full info feedback

  set $y_{t+1} \leftarrow y_t + \gamma_t \hat{v}_t$ # update scores

end for

Regret analysis

- Use constant $\gamma_t \equiv \gamma$ # complications otherwise
- Fix benchmark strategy $p \in \mathcal{X}$ and consider the ***Fenchel coupling:***

$\leq \Phi(y_t)$

$$F_t = F(p, y_t) = \sum_{\alpha \in \mathcal{A}} p_\alpha \log p_\alpha + \boxed{\log \sum_{\alpha \in \mathcal{A}} \exp(y_{\alpha,t})} - \langle y_t, p \rangle$$

- ***Energy inequality:***

$$F_{t+1} \leq F_t + \gamma \langle \hat{v}_t, x_t - p \rangle + \tfrac{1}{2}\gamma^2 \|\hat{v}_t\|_\infty^2$$

- Expand and rearrange:

$$\langle v_t, p - x_t \rangle \leq \frac{F_t - F_{t+1}}{\gamma} + \langle U_t, x_t - p \rangle + \langle b_t, x_t - p \rangle + \frac{\gamma}{2} \|\hat{v}_t\|_\infty^2$$

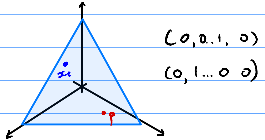
- **How to proceed?**

Regret analysis, cont'd

Bound each term separately:

- $\mathbb{E}[\langle U_t, x_t - p\rangle] = \mathbb{E}[\mathbb{E}[\langle U_t, x_t - p\rangle \mid \mathcal{F}_t]] = \mathbb{E}[\langle \cancelto{0}{\mathbb{E}[U_t \mid \mathcal{F}_t]}, x_t - p\rangle] = 0$
- $\mathbb{E}[\langle b_t, x_t - p\rangle] \le \mathbb{E}[\|b_t\|_\infty \|x_t - p\|_1] \le 2B_t$
- $\mathbb{E}[\|\hat{v}_t\|_\infty^2] \le M_t^2$

ORACLE MODEL HYPOTHESES



$\Rightarrow$ After telescoping, we get:

$$\boxed{\mathbb{E}[\mathrm{Reg}_p(T)] \le \frac{F_1}{\gamma} + 2\sum_{t=1}^{T} B_t + \frac{\gamma}{2}\sum_{t=1}^{T} M_t^2}$$

<u>NB1:</u> $\|x_t - p\|_1 \le$ diameter of simplex $= 2$

<u>NB2:</u> $F(y) = \sum_\alpha p_\alpha \log p_\alpha + \overbrace{\log \sum_\alpha \exp(y_\alpha)}^{\Phi(y)} - \langle y, p\rangle$

Minimum of $\Phi(y) - \langle y, p\rangle \rightsquigarrow \dfrac{\exp(y_\alpha)}{\underbrace{\sum_\beta \exp(y_\beta)}_{Z}} = p_\alpha \rightsquigarrow y_\alpha = \log p_\alpha - \log Z \Rightarrow \min = -\sum_\alpha p_\alpha \log p_\alpha \Rightarrow \underline{F(y) \ge 0}$

Regret analysis, cont'd

$$\boxed{\mathbb{E}[\mathrm{Reg}_p(T)] \leq \frac{F_1}{\gamma} + 2\sum_{t=1}^{T} B_t + \frac{\gamma}{2}\sum_{t=1}^{T} M_t^2}$$

- How do we get no regret?
- Need to minimize a function of the form $\frac{A}{\gamma} + B\gamma \to -\frac{A}{\gamma^2} + B = 0 \Rightarrow \gamma^2 = \frac{A}{B} \Rightarrow \gamma = \sqrt{\frac{A}{B}} = \sqrt{\frac{F_1}{2\sum_{t=1}^T M_t^2}}$

  $\hookrightarrow$ minimum value $= \dots = 2\sqrt{AB} = 2\sqrt{F_1 \cdot \frac{1}{2}\sum_{t=1}^T M_t^2} = \sqrt{2F_1 \cdot \sum_{t=1}^T M_t^2}$

Regret of Hedge-O

Theorem

☞ *Assume:*

- *Sequence of payoff vectors $v_t \in \mathbb{R}^{\mathcal{A}}$; SFO feedback* ⇝ Initialization with $y_1 = 0 \Rightarrow F_1 = \log|\mathcal{A}|$
- $\gamma = \sqrt{\dfrac{2\log m}{\sum_{t=1}^{T} M_t^2}}$

☞ *Then: for all $p \in \mathcal{X}$,* Hedge-O *enjoys the bound*

$$\mathrm{Reg}_p(T) \leq 2\sum_{t=1}^{T} B_t + \sqrt{2\log m \cdot \sum_{t=1}^{T} M_t^2}$$

Regret of Hedge-O

Theorem

☞ ***Assume:***

- *Sequence of payoff vectors* $v_t \in \mathbb{R}^{\mathcal{A}}$; *SFO feedback*
- $\gamma = \sqrt{\dfrac{2\log m}{\sum_{t=1}^{T} M_t^2}}$

☞ ***Then:*** *for all* $p \in \mathcal{X}$, Hedge-O *enjoys the bound*

$$\mathrm{Reg}_p(T) \leq 2\sum_{t=1}^{T} B_t + \sqrt{2\log m \cdot \sum_{t=1}^{T} M_t^2}$$

Remarks:

- $\mathcal{O}(\sqrt{T})$ regret if feedback is unbiased ($b_t = 0$) and has finite variance ($M_t \leq M$)
- This bound is tight in T — Abernethy et al., 2008
- Logarithmic dependence on m — Can deal with exponentially many arms!

Regret of Hedge

Theorem (Auer et al., 1995)

- ***Assume:***
  - *Sequence of payoff vectors* $v_t \in [0,1]^{\mathcal{A}}$; *Full info feedback*
  - $\gamma = \sqrt{(2\log m)/T}$
- ***Then:*** Hedge *enjoys the bound*

$$\operatorname{Reg}_p(T) \leq \sqrt{2\log m \cdot T} = \mathcal{O}(\sqrt{T})$$

Regret of Hedge

Theorem (Auer et al., 1995)

Assume:

- *Sequence of payoff vectors* $v_t \in [0,1]^{\mathcal{A}}$; *Full info feedback*
- $\gamma = \sqrt{(2\log m)/T}$

Then: Hedge *enjoys the bound*

$$\mathrm{Reg}_p(T) \le \sqrt{2\log m \cdot T} = \mathcal{O}(\sqrt{T})$$

Remarks:

- Cannot achieve $\mathcal{O}(1)$ regret as in continuous time — # Why?
- This bound is tight in T — Abernethy et al., 2008
- Logarithmic dependence on m — Can deal with exponentially many arms!

Outline

1. Online learning in continuous time
2. Online learning in discrete time
3. Learning with oracle feedback
4. Learning with bandit feedback

Learning with bandit feedback

Three types of feedback (from best to worst):

- **Full, exact information:** observe entire payoff vector v_t
- **Full, inexact information:** observe noisy estimate of v_t
- **Partial information / Bandit:** only chosen component $u_t(\alpha_t) = v_{\alpha_t,t}$

Importance weighted estimators

Fix a payoff vector $v \in \mathbb{R}^{\mathcal{A}}$ and a probability distribution P on $\mathcal{A}$. Then the ***importance weighted estimator*** of v_α is the random variable

mixed strategy

$$\hat{v}_\alpha = \frac{\mathbb{1}_\alpha}{P_\alpha} v_\alpha = \begin{cases} v_\alpha / P_\alpha & \text{if } \alpha \text{ is drawn } (\alpha = \beta) \\ 0 & \text{otherwise} \quad (\alpha \neq \beta) \end{cases} \tag{IWE}$$

IWE as an oracle model

- ***Unbiased:*** $\mathbb{E}[\hat{v}_\alpha] = v_\alpha$ $\rightsquigarrow \mathbb{E}[\hat{v}_\alpha] = \sum_\beta \cancel{P_\beta} \frac{\mathbb{1}(\alpha=\beta)}{\cancel{P_\beta}} v_\beta = \sum_\beta \mathbb{1}(\alpha=\beta) v_\beta = v_\alpha$ ☞ $b_t = 0$
- ***Second moment:*** $\mathbb{E}[\hat{v}_\alpha^2] = v_\alpha^2 / P_\alpha$ [Exercise] ☞ $M_t = \mathcal{O}(1/\min_\alpha x_{\alpha,t})$

Regret analysis, cont'd

Example:

→ Encounter but do not observe payoff vector $v_t = \begin{pmatrix} 3.2 \\ 2.1 \\ 4.8 \\ 1.9 \end{pmatrix}$

→ Play mixed strategy $x_t = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.4 \\ 0.3 \end{pmatrix}$

→ Draw arm $\alpha_t = 2$

→ Receive payoff $v_{t,\alpha_t} = 2.1$ ⚠ This is the only information we have

→ IWG estimator : $\hat{v}_t = \begin{pmatrix} 0 \\ 2.1/0.2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 10.2 \\ 0 \\ 0 \end{pmatrix}$

The EXP3 algorithm

Algorithm Exponential weights for exploration and exploitation (EXP3) # HEDGE with bandit feedback

Require: set of actions $\mathcal{A}$; sequence of payoff vectors $v_t \in [0,1]^{\mathcal{A}}$, $t = 1, 2, \ldots$

Initialize: $y_1 \in \mathbb{R}^{\mathcal{A}}$

for all $t = 1, 2, \ldots$ **do**

set $x_t \leftarrow \Lambda(y_t)$ # mixed strategy

play $\alpha_t \sim x_t$ and **receive** $v_{\alpha_t,t}$ # choose action / get payoff

set $\hat{v}_t \leftarrow \dfrac{v_{\alpha_t,t}}{x_{\alpha_t,t}} e_{\alpha_t}$ # IW estimator

set $y_{t+1} \leftarrow y_t + \gamma_t \hat{v}_t$ # update scores

end for

Regret analysis

- Use constant $\gamma_t \equiv \gamma$ # complications otherwise
- Fix benchmark strategy $p \in \mathcal{X}$ and consider the ***Fenchel coupling:***

$$F_t = F(p, y_t) = \sum_{\alpha \in \mathcal{A}} p_\alpha \log p_\alpha + \log \sum_{\alpha \in \mathcal{A}} \exp(y_{\alpha,t}) - \langle y_t, p \rangle$$

- ***Energy inequality:***

$$F_{t+1} \le F_t + \gamma \langle \hat{v}_t, x_t - p \rangle + \tfrac{1}{2} \gamma^2 \|\hat{v}_t\|_\infty^2$$

- Expand and rearrange:

$$\langle v_t, p - x_t \rangle \le \frac{F_t - F_{t+1}}{\gamma} + \langle U_t, x_t - p \rangle + \frac{\gamma}{2} \|\hat{v}_t\|_\infty^2$$

- No bias, **but** $\mathbb{E}[\|\hat{v}_t\|_\infty^2] = \mathcal{O}(1/\min_\alpha x_{\alpha,t})$ **is unbounded ✗**
- **How to proceed?**

$\cancel{\le M_t^2}$

Energy inequality

Basic lemma

Fix some $y, w \in \mathbb{R}^{\mathcal{A}}$, and let $x \propto \exp(y)$. Then:

$$\log \sum_{\alpha \in \mathcal{A}} \exp(y_\alpha + w_\alpha) \leq \log \sum_{\alpha \in \mathcal{A}} \exp(y_\alpha) + \langle x, w \rangle + \tfrac{1}{2} \|w\|_\infty^2$$

Energy inequality

Basic lemma

Fix some $y \in \mathbb{R}^{\mathcal{A}}$, $w \in (-\infty, 1]^{\mathcal{A}}$, and let $x \propto \exp(y)$. Then:

$$\log \sum_{\alpha \in \mathcal{A}} \exp(y_\alpha + w_\alpha) \leq \log \sum_{\alpha \in \mathcal{A}} \exp(y_\alpha) + \langle x, w \rangle + \sum_{\alpha \in \mathcal{A}} x_\alpha w_\alpha^2$$

Proof.

- Key element of the proof: if $t \leq 1$, $e^t \leq 1 + t + t^2$

$\square$

Regret analysis, cont'd

- Refined energy inequality: $F_{t+1} \le F_t + \gamma \langle \hat{v}_t, x_t - p \rangle + \frac{\gamma^2}{2} \sum_\alpha x_{\alpha,t} \hat{v}_{\alpha,t}^2$

$$\Rightarrow \quad \langle \hat{v}_t, p - x_t \rangle \le \frac{F_t - F_{t+1}}{\gamma} + \frac{\gamma}{2} \sum_\alpha x_{\alpha,t} \hat{v}_{\alpha,t}^2$$

- Average + Telescope

$$\underbrace{\sum_{t=1}^T \mathbb{E}[\langle \hat{v}_t, p - x_t \rangle]}_{\mathbb{E}[\mathrm{Reg}_p(T)]} \le \frac{F_1}{\gamma} + \frac{\gamma}{2} \sum_{t=1}^T \sum_\alpha \mathbb{E}[x_{\alpha,t} \hat{v}_{\alpha,t}^2]$$

$$\le \frac{F_1}{\gamma} + \frac{\gamma}{2} \sum_{t=1}^T \sum_{\alpha,\beta} x_{\alpha,t} \cancel{x_{\beta,t}} \frac{\mathbb{1}(\alpha=\beta)}{x_{\beta,t}^{\cancel{2}}} v_{\beta,t}^2$$

$$= \frac{F_1}{\gamma} + \frac{\gamma}{2} \sum_{t=1}^T \sum_\alpha \cancel{x_{\alpha,t}} \frac{1}{\cancel{x_{\alpha,t}}} \underbrace{v_{\alpha,t}^2}_{\le 1}$$

$$\le \frac{F_1}{\gamma} + \frac{\gamma}{2} \sum_{t=1}^T \sum_\alpha 1 \le \frac{F_1}{\gamma} + \frac{\gamma A T}{2}$$

Regret of EXP3

Theorem (Auer et al., 1995)

☞ ***Assume:***

- *EXP3 is run for T iterations with $\gamma = \sqrt{\log m/(mT)}$*

- ***Then:*** *For all $p \in \mathcal{X}$, the learner enjoys the bound*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq 2\sqrt{m \log m \cdot T}$$

Regret of EXP3

Theorem (Auer et al., 1995)

☞ ***Assume:***

- *EXP3 is run for T iterations with $\gamma = \sqrt{\log m/(mT)}$*

- ***Then:*** *For all $p \in \mathcal{X}$, the learner enjoys the bound*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq 2\sqrt{m \log m \cdot T}$$

Remarks:

✓ Tight in T ➭ Abernethy et al., 2008

✗ Worse than full info bound by a factor of $\sqrt{m}$ # cf. Hedge-O

- Regret can be improved to $\mathcal{O}(\sqrt{mT})$ **but no lower** ➭ Audibert & Bubeck, 2010; Abernethy et al., 2015

- T must be known ⚠ Thoughts?

References I

[1] Abernethy, J., Bartlett, P. L., Rakhlin, A., and Tewari, A. Optimal strategies and minimax lower bounds for online convex games. In *COLT '08: Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

[2] Abernethy, J., Lee, C., and Tewari, A. Fighting bandits with a new kind of smoothness. In *NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015.

[3] Audibert, J.-Y. and Bubeck, S. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11: 2635-2686, 2010.

[4] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

[5] Blackwell, D. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1-8, 1956.

[6] Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1-122, 2012.

[7] Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[8] Fudenberg, D. and Levine, D. K. *The Theory of Learning in Games*, volume 2 of *Economic learning and social evolution*. MIT Press, Cambridge, MA, 1998.

[9] Hannan, J. Approximation to Bayes risk in repeated play. In Dresher, M., Tucker, A. W., and Wolfe, P. (eds.), *Contributions to the Theory of Games, Volume III*, volume 39 of *Annals of Mathematics Studies*, pp. 97-139. Princeton University Press, Princeton, NJ, 1957.

[10] Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107-194, 2011.

[11] Sorin, S. Exponential weight algorithm in continuous time. *Mathematical Programming*, 116(1):513-528, 2009.