ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΠΑΙΓΝΙΩΝ ΚΑΙ ΛΗΨΗΣ ΑΠΟΦΑΣΕΩΝ

ΛΗΣΤΕΣ, ΚΟΥΛΟΧΕΡΗΔΕΣ, ΚΑΙ ΘΕΩΡΙΑ ΜΑΘΗΣΗΣ

Παναγιώτης Μερτικόπουλος

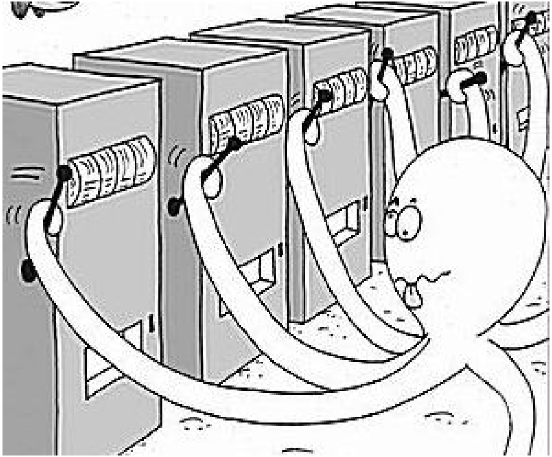
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Τμήμα Μαθηματικών

Χειμερινό Εξάμηνο, 2023–2024

Multi-armed bandits

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



Outline

Game-theoretic learning

Sequence of events – continuous time

Require: finite game $\Gamma \equiv \Gamma(\mathcal{N}, \mathcal{A}, u)$

repeat

At each epoch $t \geq 0$ **do simultaneously** for all players $i \in \mathcal{N}$ # continuous time

Choose ***mixed strategy*** $x_i(t) \in \mathcal{X}_i := \Delta(\mathcal{A}_i)$ # mixing

Encounter ***mixed payoff vector*** $v_i(x(t))$ and get **mixed payoff** $u_i(x(t)) = \langle v_i(t), x(t) \rangle$ # feedback phase

until end

Defining elements

- **Time:** $t \geq 0$
- **Players:** finite
- **Actions:** finite
- **Payoffs:** game
- **Feedback:** mixed payoff vectors

Online learning

Sequence of events – continuous time

Require: set of actions $\mathcal{A} = \{1, \dots, m\}$, stream of payoff vectors $v_t \in [0,1]^{\mathcal{A}}$, $t \geq 0$

repeat

At each epoch $t \geq 0$ **do** # continuous time

Choose ***mixed strategy*** $x_t \in \mathcal{X}$ # mixing

Encounter ***payoff vector*** v_t and get ***mixed payoff*** $u_t(x_t) = \langle v_t, x_t \rangle$ # feedback phase

until end

Defining elements

- **Time:** $t \geq 0$
- **Players:** ***single*** # *"unilateral viewpoint"*
- **Actions:** finite
- **Payoffs:** ***exogenous*** # *"game against Nature"*
- **Feedback:** mixed payoff vectors

Online v. multi-agent learning

How are payoffs generated?

- Multi-agent viewpoint
  - *Multiple agents*
  - *Endogenous rewards:* individual payoffs depend on other agents
  - *Game-theoretic:* underlying mechanism is a (finite) game
- Online viewpoint
  - *Single agent*
  - *Exogenous rewards:* different payoff vector at each stage
  - *Agnostic:* no assumptions on mechanism generating $v(t)$ # dispassionate Nature

Online v. multi-agent learning

How are payoffs generated?

- Multi-agent viewpoint
  - *Multiple agents*
  - *Endogenous rewards:* individual payoffs depend on other agents
  - *Game-theoretic:* underlying mechanism is a (finite) game
- Online viewpoint
  - *Single agent*
  - *Exogenous rewards:* different payoff vector at each stage
  - *Agnostic:* no assumptions on mechanism generating $v(t)$ # dispassionate Nature

What is the interplay between online and multi-agent learning?

The agent's regret

Performance of a policy x_t measured by the agent's **regret**

$$u_t(p) - u_t(x_t)$$

The agent's regret

Performance of a policy x_t measured by the agent's **regret**

$$\int_0^T [u_t(p) - u_t(x_t)]\, dt$$

The agent's regret

Performance of a policy x_t measured by the agent's **regret**

$$\max_{p \in \mathcal{X}} \int_0^T [u_t(p) - u_t(x_t)] \, dt$$

The agent's regret

Performance of a policy x_t measured by the agent's **regret**

$$\mathrm{Reg}(T) = \max_{p \in \mathcal{X}} \int_0^T [u_t(p) - u_t(x_t)]\, dt = \max_{p \in \mathcal{X}} \int_0^T \langle v_t, p - x_t \rangle\, dt$$

The agent's regret

Performance of a policy x_t measured by the agent's **regret**

$$\mathrm{Reg}(T) = \max_{p \in \mathcal{X}} \int_0^T [u_t(p) - u_t(x_t)]\, dt = \max_{p \in \mathcal{X}} \int_0^T \langle v_t, p - x_t \rangle\, dt$$

No regret: $\mathrm{Reg}(T) = o(T)$ # the smaller the better

"The chosen policy is as good as the best fixed strategy in hindsight."

The agent's regret

Performance of a policy x_t measured by the agent's **regret**

$$\operatorname{Reg}(T) = \max_{p \in \mathcal{X}} \int_0^T [u_t(p) - u_t(x_t)]\, dt = \max_{p \in \mathcal{X}} \int_0^T \langle v_t, p - x_t \rangle\, dt$$

No regret: $\operatorname{Reg}(T) = o(T)$ # the smaller the better

"The chosen policy is as good as the best fixed strategy in hindsight."

Prolific literature:

- Economics ➽ Hannan (1957), Fudenberg & Levine (1998)
- Mathematics ➽ Blackwell (1956), Bubeck & Cesa-Bianchi (2012)
- Computer science ➽ Shalev-Shwartz (2011), Cesa-Bianchi & Lugosi (2006)

Exponential weights for online learning

Exponential weight dynamics

$$\dot{y}_t = v_t \qquad x_t = \Lambda(y_t) \tag{EWD}$$

where $\Lambda : \mathbb{R}^{\mathcal{A}} \to \mathcal{X}$ is the ***logit map***

$$\Lambda_\alpha(y) = \frac{\exp(y_\alpha)}{\sum_{\beta \in \mathcal{A}} \exp(y_\beta)}$$

Does (EWD) lead to no regret?

Bounding the regret

- Fix a comparator $p \in \mathcal{X}$
- Consider associated regret

$$\operatorname{Reg}_p(T) = \int_0^T \langle v_t, p - x_t \rangle \, dt$$

Bounding the regret

- Fix a comparator $p \in \mathcal{X}$
- Consider associated regret

$$\mathrm{Reg}_p(T) = \int_0^T \langle v_t, p - x_t \rangle \, dt$$

- Focus on integrand

$$\langle v_t, x_t - p \rangle = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

Bounding the regret

- Fix a comparator $p \in \mathcal{X}$
- Consider associated regret

$$\mathrm{Reg}_p(T) = \int_0^T \langle v_t, p - x_t \rangle \, dt$$

- Focus on integrand

$$\langle v_t, x_t - p \rangle = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

- Suppose we can find a ***potential function*** $\Phi(y)$ such that

$$\nabla \Phi(y) = \Lambda(y) - p \implies \frac{d\Phi}{dt} = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

Bounding the regret

- Fix a comparator $p \in \mathcal{X}$
- Consider associated regret

$$\mathrm{Reg}_p(T) = \int_0^T \langle v_t, p - x_t \rangle \, dt$$

- Focus on integrand

$$\langle v_t, x_t - p \rangle = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

- Suppose we can find a ***potential function*** $\Phi(y)$ such that

$$\nabla \Phi(y) = \Lambda(y) - p \implies \frac{d\Phi}{dt} = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

- Then

$$\mathrm{Reg}_p(T) = -\int_0^T \frac{d\Phi}{dt} \, dt = \Phi(y_0) - \Phi(y_T)$$

Bounding the regret

- Fix a comparator $p \in \mathcal{X}$
- Consider associated regret

$$\mathrm{Reg}_p(T) = \int_0^T \langle v_t, p - x_t \rangle \, dt$$

- Focus on integrand

$$\langle v_t, x_t - p \rangle = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

- Suppose we can find a ***potential function*** $\Phi(y)$ such that

$$\nabla \Phi(y) = \Lambda(y) - p \implies \frac{d\Phi}{dt} = \langle \dot{y}_t, \Lambda(y_t) - p \rangle$$

- Then

$$\mathrm{Reg}_p(T) = -\int_0^T \frac{d\Phi}{dt} \, dt = \Phi(y_0) - \Phi(y_T)$$

If suitable potential exists $\implies \mathrm{Reg}(T) \leq \Phi(y_0) - \min \Phi$

Finding a potential

What could a potential function look like?

Minimizing the potential

What is the minimum value of the potential?

Energy functions

We can encode the above with the help of the following ***energy functions:***

- **The Fenchel coupling:**

$$F(p, y) = \sum_{\alpha \in \mathcal{A}} p_\alpha \log p_\alpha + \log \sum_{\alpha \in \mathcal{A}} \exp(y_\alpha) - \sum_{\alpha \in \mathcal{A}} p_\alpha y_\alpha$$

- Substituting $x \leftarrow \Lambda(y)$ yields the **Kullback-Leibler divergence:**

$$D_{\mathrm{KL}}(p, x) = \sum_{\alpha \in \mathcal{A}} p_\alpha \log \frac{p_\alpha}{x_\alpha}$$

Key property: $\dfrac{d}{dt} F(p, y_t) = \langle v_t, x_t - p \rangle$

Regret of (EWD)

Theorem (Sorin, 2009)

Under (EWD), *the learner enjoys the regret bound*

$$\mathrm{Reg}_p(T) \leq F(p, y_0) = \sum_{\alpha \in \mathcal{A}} p_\alpha \log p_\alpha + \log \sum_{\alpha \in \mathcal{A}} \exp(y_{\alpha,0}) - \sum_{\alpha \in \mathcal{A}} p_\alpha y_{\alpha,0}$$

In particular, if (EWD) *is initialized with* $y_0 = 0$, *we have*

$$\mathrm{Reg}(T) \leq \log m$$

Outline

Online learning in discrete time

Sequence of events – discrete time

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

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

Choose **mixed strategy** $x_t \in \mathcal{X} := \Delta(\mathcal{A})$

Play **action** $\alpha_t \sim x_t$

Encounter **payoff vector** v_t and receive **payoff** $u_t(\alpha_t) = v_{\alpha_t,t}$

end for

Defining elements

- Time: ***discrete***
- Players: single
- Actions: finite
- Payoffs: exogenous
- Feedback: ***depends*** (full or partial information, ...)

Online learning in discrete time

Sequence of events – discrete time

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

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

Choose mixed strategy $x_t \in \mathcal{X} := \Delta(\mathcal{A})$

Play action $\alpha_t \sim x_t$

Encounter payoff vector v_t and receive payoff $u_t(\alpha_t) = v_{\alpha_t,t}$

end for

Regret

$$\mathrm{Reg}(T) = \max_{p \in \mathcal{X}} \sum_{t=1}^{T} \Big[\mathbb{E}_{v_{\alpha_t,t}}[\alpha_t \sim p] - \mathbb{E}_{v_{\alpha_t,t}}[\alpha_t \sim x_t]\Big] = \max_{p \in \mathcal{X}} \sum_{t=1}^{T} \langle v_t, p - x_t \rangle$$

The feedback process

Types of feedback

From best to worst (more to less info):

- ***Full information:*** v_t # deterministic vector feedback
- ***Noisy payoff vectors:*** $v_t + Z_t$ # stochastic vector feedback
- ***Bandit / Payoff-based:*** $u_t(\alpha_t) = v_{\alpha_t,t}$ # stochastic scalar feedback

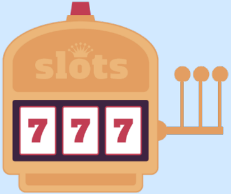
The feedback process

Types of feedback

From best to worst (more to less info):

- ***Full information:*** v_t # deterministic vector feedback
- ***Noisy payoff vectors:*** $v_t + Z_t$ # stochastic vector feedback
- ***Bandit / Payoff-based:*** $u_t(\alpha_t) = v_{\alpha_t,t}$ # stochastic scalar feedback

Example



Play $x_t \leftarrow (1/2, 1/3, 1/6)$ $\rightsquigarrow$ Draw $\alpha_t \leftarrow 1$

Full information

v_t 1 3 2

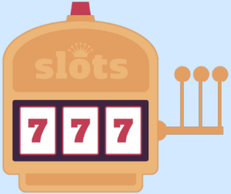
The feedback process

Types of feedback

From best to worst (more to less info):

- ***Full information:*** v_t # deterministic vector feedback
- ***Noisy payoff vectors:*** $v_t + Z_t$ # stochastic vector feedback
- ***Bandit / Payoff-based:*** $u_t(\alpha_t) = v_{\alpha_t, t}$ # stochastic scalar feedback

Example



Play $x_t \leftarrow (1/2, 1/3, 1/6)$ $\rightsquigarrow$ Draw $\alpha_t \leftarrow 1$

Noisy payoff vectors

$v_t + Z_t$ 1.4 2.9 1.2

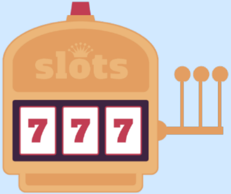
The feedback process

Types of feedback

From best to worst (more to less info):

- ***Full information:*** v_t # deterministic vector feedback
- ***Noisy payoff vectors:*** $v_t + Z_t$ # stochastic vector feedback
- ***Bandit / Payoff-based:*** $u_t(\alpha_t) = v_{\alpha_t,t}$ # stochastic scalar feedback

Example



Play $x_t \leftarrow (1/2, 1/3, 1/6)$ $\rightsquigarrow$ Draw $\alpha_t \leftarrow 1$

Bandit / Payoff-based

$v_{\alpha_t,t}$ 1 ✗ ✗

The feedback process

Types of feedback

From best to worst (more to less info):

- ***Full information:*** v_t # deterministic vector feedback
- ***Noisy payoff vectors:*** $v_t + Z_t$ # stochastic vector feedback
- ***Bandit / Payoff-based:*** $u_t(\alpha_t) = v_{\alpha_t,t}$ # stochastic scalar feedback

Defining features:

- ***Vector*** (all payoffs) vs. ***Scalar*** (bandit)
- ***Deterministic*** (full info) vs. ***Stochastic*** (noisy, bandit)

☞ Randomness defined relative to ***history of play*** $\mathcal{F}_t := \mathcal{F}(x_1, \dots, x_t)$

☞ Other feedback models also possible (noisy / delayed observations,...)

Regret

The agent's **regret** in discrete time

$$\textbf{Realized regret:}\quad \operatorname{Reg}(T) = \max_{\alpha\in\mathcal{A}} \sum_{t=1}^{T} [u_t(\alpha) - u_t(\alpha_t)]$$

$$\textbf{Mean regret:}\quad \overline{\operatorname{Reg}}(T) = \max_{p\in\mathcal{X}} \sum_{t=1}^{T} [u_t(p) - u_t(x_t)] = \max_{p\in\mathcal{X}} \sum_{t=1}^{T} \langle v_t, p - x_t\rangle$$

Regret

The agent's **regret** in discrete time

Realized regret: $$\operatorname{Reg}(T) = \max_{\alpha \in \mathcal{A}} \sum_{t=1}^{T} [u_t(\alpha) - u_t(\alpha_t)]$$

Mean regret: $$\overline{\operatorname{Reg}}(T) = \max_{p \in \mathcal{X}} \sum_{t=1}^{T} [u_t(p) - u_t(x_t)] = \max_{p \in \mathcal{X}} \sum_{t=1}^{T} \langle v_t, p - x_t \rangle$$

- **Adversarial framework:** regret guarantees against *any* given sequence v_t
- No distinction between ***mean*** regret and ***pseudo***-regret

➥ Bubeck & Cesa-Bianchi (2012)

- **Not here:** stochastic, Markovian, oblivious/non-oblivious,...

➥ Cesa-Bianchi & Lugosi (2006)

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 $u_t(\alpha_t) = v_{\alpha_t,t}$

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 $u_t(\alpha_t) = v_{\alpha_t,t}$

The oracle model

A ***stochastic first-order oracle (SFO)*** model of v_t is a random vector 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:*** $\|b_t\| \leq B_t$
- ***Variance:*** $\mathbb{E}[\|U_t\|^2 \mid \mathcal{F}_t] \leq \sigma_t^2$
- ***Second moment:*** $\mathbb{E}[\|\hat{v}_t\|^2 \mid \mathcal{F}_t] \leq M_t^2$

Reconstructing payoff vectors

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 relative to P is the random variable

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

IWE as an oracle model

- ***Unbiased:***

$$\mathbb{E}[\hat{v}_\alpha] = v_\alpha$$

- ***Second moment:***

$$\mathbb{E}[\hat{v}_\alpha^2] = \frac{v_\alpha^2}{P_\alpha}$$

The Hedge algorithm

Algorithm HEDGE — # EXPWEIGHT with full information

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
- **observe** v_t — # full info feedback
- **set** $y_{t+1} \leftarrow y_t + \gamma_t v_t$ — # update scores

end for

Basic idea:

- Aggregate payoff information
- Choose actions with probability exponentially proportional to their scores
- Rinse & repeat

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:***

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

- Telescope to get

$$\boxed{\mathrm{Reg}_p(T) \leq \frac{F_1}{\gamma} + \frac{\gamma T}{2}}$$

- **How to proceed?**

Regret analysis, cont'd

How to choose γ?

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*

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

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*

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

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

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 $v_t \in \mathbb{R}^{\mathcal{A}}$, $t = 1, 2, \dots$

- **Initialize:** $y_1 \in \mathbb{R}^{\mathcal{A}}$
- **for all** $t = 1, 2, \dots$ **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:***

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

- ***Energy inequality:***

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

- 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:

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*

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

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*

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

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*

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

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

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

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

IWE as an oracle model

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

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 $v_t \in [0,1]^{\mathcal{A}}$, $t = 1, 2, \dots$

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

for all $t = 1, 2, \dots$ **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:***

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

- Expand and rearrange:

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

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

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.

$\square$

Regret analysis, cont'd

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.