Learning to Play Games Using Multiplicative Weights

Rob Schapire

including joint work with:
Peter Auer
Nicolò Cesa-Bianchi
Yoav Freund

www.research.att.com/∼schapire

Outline

Part I: Playing Repeated Games [with Freund]

- quick review of game theory
- multiplicative-weights algorithm and analysis
- simple proof of minmax theorem
- method of approximately solving a game
- alternate goal for learning
  - analysis and very tight lower bound

Part II: Learning with Restricted Feedback
(or: The Adversarial Multi-armed Bandit Problem)

[with Auer, Cesa-Bianchi & Freund]

Games

- game defined by matrix $\mathbf{M}$:

	Rock	Paper	Scissors
Rock	$1/2$	1	0
Paper	0	$1/2$	1
Scissors	1	0	$1/2$

- row player ("learner") chooses row i
- column player ("environment") chooses column j (simultaneously)
- learner's goal: minimize loss $\mathbf{M}(i, j)$

Randomized Play

- usually allow randomized play:
  - learner chooses distribution $\mathbf{P}$ over rows
  - environment chooses distribution $\mathbf{Q}$ over columns
- learner's (expected) loss

$$\begin{aligned} &= \sum_{i,j} \mathbf{P}(i)\mathbf{M}(i,j)\mathbf{Q}(j) \\ &= \mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{Q} \equiv \mathbf{M}(\mathbf{P},\mathbf{Q}) \end{aligned}$$

- i, j = "pure" strategies
- $\mathbf{P}$, $\mathbf{Q}$ = "mixed" strategies
- n = # rows of $\mathbf{M}$

The Minmax Theorem

- von Neumann's minmax theorem:

$$\begin{aligned}\min_{\mathbf{P}} \max_{\mathbf{Q}} \mathbf{M}(\mathbf{P},\mathbf{Q}) &= \max_{\mathbf{Q}} \min_{\mathbf{P}} \mathbf{M}(\mathbf{P},\mathbf{Q}) \\ &= v \\ &= \text{“value of game } \mathbf{M}\end{aligned}$$

- in words:
  - $v = \min \max$ means:
    - learner has strategy $\mathbf{P}^*$ such that $\forall$ environment strategy $\mathbf{Q}$ loss $\mathbf{M}(\mathbf{P}^*,\mathbf{Q}) \le v$
  - $v = \max \min$ means:
    - this is optimal in sense that environment has strategy $\mathbf{Q}^*$ such that $\forall$ learner strategy $\mathbf{P}$ loss $\mathbf{M}(\mathbf{P},\mathbf{Q}^*) \ge v$

Minmax Play

- classical game theory says:
  - given game **M**, should play using minmax strategy
  - (can compute with linear programming)
- problems:
  - **M** may be unknown
  - **M** may be too large to compute minmax strategy
  - environment may not be adversary (may admit loss $\ll v$)
- hopeless for one-shot game
- but: for repeated games, can one learn to play well?

Repeated Play

- $\mathbf{M}$ unknown
- for $t = 1, \ldots, T$:
  - learner chooses $\mathbf{P}_t$
  - environment chooses $\mathbf{Q}_t$ (may depend on $\mathbf{P}_t$)
  - learner observes loss of each pure strategy i: $\mathbf{M}(i, \mathbf{Q}_t)$
  - learner's loss $= \mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)$
- goal: do almost as well as best strategy against actual sequence of plays $\mathbf{Q}_1, \ldots, \mathbf{Q}_T$
- want:

$$\underbrace{\sum_{t=1}^{T} \mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)}_{\text{learner's loss}} \leq \underbrace{\min_{\mathbf{P}} \sum_{t=1}^{T} \mathbf{M}(\mathbf{P}, \mathbf{Q}_t)}_{\text{“best” loss}} + [\text{“small amount”}]$$

Algorithm MW

- direct generalization of Littlestone & Warmuth's "weighted majority algorithm"
  - independently discovered by [Fudenberg & Levine '95]
- learning rate $\eta > 0$
- initially, $\mathbf{P}_1(i) = 1/n$ (uniform)
- after round t, update:

$$\mathbf{P}_{t+1}(i) = \frac{\mathbf{P}_t(i)\exp(-\eta \mathbf{M}(i, \mathbf{Q}_t))}{Z_t}$$

where Z_t is normalization factor

Main Theorem

- Theorem: for any $\mathbf{M}$ with entries in $[0,1]$
  for any environment

$$\sum_{t=1}^{T} \mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t) \le a_\eta \min_{\mathbf{P}} \sum_{t=1}^{T} \mathbf{M}(\mathbf{P}, \mathbf{Q}_t) + c_\eta \ln n$$

  where

$$a_\eta = \frac{\eta}{1 - e^{-\eta}} \qquad c_\eta = \frac{1}{1 - e^{-\eta}}$$

- e.g.:

η	a_η	c_η
0.1	1.05	10.5
1	1.58	1.58
5	5.04	1.01

Corollary

- Corollary: If T known, can choose η so that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{M}(\mathbf{P}_t,\mathbf{Q}_t) \le \min_{\mathbf{P}}\frac{1}{T}\sum_{t=1}^{T}\mathbf{M}(\mathbf{P},\mathbf{Q}_t) + \Delta_T$$

  where $\Delta_T = O\left(\sqrt{\frac{\ln n}{T}}\right)$
- relative loss bound
- very moderate in n (number of rows)
- results of this type first proved by [Hannan '57] and [Blackwell '56]
  - and later by [Foster & Vohra '93]

Proof Sketch

- direct generalization of Littlestone & Warmuth
- based on relative entropy (a.k.a. Kullback-Liebler divergence):

$$\mathrm{RE}\left(\mathbf{P}_1 \parallel \mathbf{P}_2\right) = \sum_i \mathbf{P}_1(i) \ln\left(\frac{\mathbf{P}_1(i)}{\mathbf{P}_2(i)}\right)$$

- Main Lemma: for any $\mathbf{P}$

$$\mathrm{RE}\left(\mathbf{P} \parallel \mathbf{P}_{t+1}\right) - \mathrm{RE}\left(\mathbf{P} \parallel \mathbf{P}_t\right) \leq \eta\mathbf{M}(\mathbf{P}, \mathbf{Q}_t) - (1 - e^{-\eta})\mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)$$

- Proof of lemma: mostly plugging in definitions
- Proof of main theorem:
  sum over $t = 1, \ldots, T$:

$$\mathrm{RE}\left(\mathbf{P} \parallel \mathbf{P}_{T+1}\right) - \mathrm{RE}\left(\mathbf{P} \parallel \mathbf{P}_1\right) \leq \eta \sum_{t=1}^{T} \mathbf{M}(\mathbf{P}, \mathbf{Q}_t) - (1 - e^{-\eta}) \sum_{t=1}^{T} \mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)$$

  and rearrange terms

Corollary

- Corollary:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{M}(\mathbf{P}_t,\mathbf{Q}_t) \le v + \Delta_T$$

- Proof:
  - let $\mathbf{P}^*$ be minmax strategy
    $\Rightarrow \forall \mathbf{Q} : \mathbf{M}(\mathbf{P}^*,\mathbf{Q}) \le v$
  - then

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbf{M}(\mathbf{P}_t,\mathbf{Q}_t) &\le \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}(\mathbf{P}^*,\mathbf{Q}_t) + \Delta_T \\ &\le v + \Delta_T\end{aligned}$$

Proof of Minmax Theorem

- want to prove:

$$\min_{\mathbf{P}} \max_{\mathbf{Q}} \mathbf{M}(\mathbf{P},\mathbf{Q}) \le \max_{\mathbf{Q}} \min_{\mathbf{P}} \mathbf{M}(\mathbf{P},\mathbf{Q})$$

  (proof of "$\ge$" is easy)

- on round t,
  environment chooses: $\mathbf{Q}_t = \arg\max_{\mathbf{Q}} \mathbf{M}(\mathbf{P}_t,\mathbf{Q})$
- let $\overline{\mathbf{P}} = \frac{1}{T}\sum_t \mathbf{P}_t \qquad \overline{\mathbf{Q}} = \frac{1}{T}\sum_t \mathbf{Q}_t$
- then

$$\begin{aligned}
\min_{\mathbf{P}} \max_{\mathbf{Q}} \mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{Q} &\le \max_{\mathbf{Q}} \overline{\mathbf{P}}^{\mathrm{T}}\mathbf{M}\mathbf{Q} \\
&= \max_{\mathbf{Q}} \frac{1}{T}\sum_t \mathbf{P}_t{}^{\mathrm{T}}\mathbf{M}\mathbf{Q} \\
&\le \frac{1}{T}\sum_t \max_{\mathbf{Q}} \mathbf{P}_t{}^{\mathrm{T}}\mathbf{M}\mathbf{Q} \\
&= \frac{1}{T}\sum_t \mathbf{P}_t{}^{\mathrm{T}}\mathbf{M}\mathbf{Q}_t \\
&\le \min_{\mathbf{P}} \frac{1}{T}\sum_t \mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{Q}_t + \Delta_T \\
&= \min_{\mathbf{P}} \mathbf{P}^{\mathrm{T}}\mathbf{M}\overline{\mathbf{Q}} + \Delta_T \\
&\le \max_{\mathbf{Q}} \min_{\mathbf{P}} \mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{Q} + \Delta_T
\end{aligned}$$

Solving a Game

- derivation shows that:

$$\max_{\mathbf{Q}} \mathbf{M}(\overline{\mathbf{P}}, \mathbf{Q}) \leq v + \Delta_T$$

  and

$$\min_{\mathbf{P}} \mathbf{M}(\mathbf{P}, \overline{\mathbf{Q}}) \geq v - \Delta_T$$

- so: $\overline{\mathbf{P}}$ and $\overline{\mathbf{Q}}$ are Δ_T-approximate minmax and maxmin strategies
- further: can choose $\mathbf{Q}_t$'s to be basis vectors

  $\Rightarrow \overline{\mathbf{Q}} = \frac{1}{T} \sum_t \mathbf{Q}_t$ sparse

  ($\leq T$ nonzero entries)

A Different Goal for Repeated Play

- say upper bound u known on value v
- different goal: limit how often loss exceeds $u + \epsilon$
- modified algorithm varies η with t:

$$\eta_t = \ln\left(\frac{(1-u)\mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)}{u(1-\mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t))}\right)$$

Analysis

- Theorem: Let $S =$ number rounds on which $\mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)$ exceeds $u + \epsilon$. Then

$$S \le \frac{\ln n}{\mathrm{RE}\left(u \ \|\ u + \epsilon\right)}$$

- Lemma: if $\max_{\mathbf{Q}} \mathbf{M}(\mathbf{P}^*, \mathbf{Q}) \le u$ then

$$\mathrm{RE}\left(\mathbf{P}^* \ \|\ \mathbf{P}_{t+1}\right) \le \mathrm{RE}\left(\mathbf{P}^* \ \|\ \mathbf{P}_t\right) - \mathrm{RE}\left(u \ \|\ \mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)\right)$$

- Proof of theorem:

$$\begin{aligned} 0 &\le \mathrm{RE}\left(\mathbf{P}^* \ \|\ \mathbf{P}_{T+1}\right) \\ &\le \mathrm{RE}\left(\mathbf{P}^* \ \|\ \mathbf{P}_1\right) - \sum_t \mathrm{RE}\left(u \ \|\ \mathbf{M}(\mathbf{P}_t, \mathbf{Q}_t)\right) \\ &\le \ln n - S\,\mathrm{RE}\left(u \ \|\ u + \epsilon\right) \end{aligned}$$

Lower Bound

- bound is extremely tight
- Theorem:

  $\forall$ u, ϵ
  $\forall$ game-playing algorithm A
  $\forall$ n sufficiently large
  $\exists$ matrix $\mathbf{M}$ and environment such that

  1. value$(\mathbf{M}) \leq u$
  2. A suffers loss $\geq u + \epsilon$ on all T rounds where

$$T = \frac{(1 - o(1)) \ln n}{\mathrm{RE}\left(u \parallel u + \epsilon\right)}$$

- proof idea:
  - choose $\mathbf{M} \in \{0, 1\}^{n \times T}$ at random
  - each entry is 1 with probability $u + \epsilon$
  - environment chooses column t on round t
  - show (1) and (2) hold with positive probability

Summary of Part I

- presented multiplicative-weights algorithm MW
  - simple, highly effective
  - performance approaches that of best strategy in hindsight
  - bounds are non-asymptotic
  - can be analyzed with no statistical assumptions
- simple proof of minmax theorem and method of approximately solving a game
- can also use (variant of) MW to limit how often loss exceeds chosen amount
  - bounds are extremely tight

Part II: Learning with Restricted Feedback

- so far, assumed learner can observe entire column selected by environment
  - i.e., can observe loss associated with all strategies
- more realistic: observe only loss for chosen strategy

Multi-Armed Bandit Problem

- K possible actions
- at each time step, learner chooses one action and gets some payoff (reward, gain)
- over time, want total gain to approach gain of best single action
- exploration versus exploitation
- traditionally, assume actions have stochastic payoffs
- this talk: allow adversary to choose payoffs

Outline

Part I: Playing Repeated Games [with Freund]

Part II: Learning with Restricted Feedback
(or: The Adversarial Multi-armed Bandit Problem)
[with Auer, Cesa-Bianchi & Freund]

- Exp3 algorithm and analysis of expected regret
  - application to repeated games
- lower bound
- analysis of actual regret
- using expert advice

Formalizing the Bandit Problem

- actions $1, \dots, K$
- for $t = 1, \dots, T$
  - learner chooses $\mathbf{p}(t)$ distribution over actions
  - environment chooses $x_i(t) \in [0, 1]$
    - $x_i(t)$ = gain if action i chosen
  - action i_t chosen randomly according to $\mathbf{p}(t)$
  - learner receives reward $x_{i_t}(t)$
- G_A = total gain of learner using algorithm A
  $= \sum_t x_{i_t}(t)$
- G_i = total gain of action i
  $= \sum_t x_i(t)$
- want G_A "close" to "best" G_i
- call this "partial information game"

Full Information Game

- learner can observe $x_i(t)$ for all i
  - i.e., case studied in first half of talk
- can use MW algorithm
- learning rate $\eta > 0$
- initially, $p_i(1) = 1/K$
- after round t, update

$$p_i(t+1) = \frac{p_i(t)\exp(\eta\, x_i(t))}{Z_t}$$

  where Z_t is a normalization factor
- can show:

$$\mathbf{E}\left[G_{\mathrm{MW}}\right] \geq \mathbf{E}\left[\max_i G_i\right] - \sqrt{2T \ln K}$$

- all expectations over random choice of actions

From Full to Partial Information Game

- want to use full information game for bandit problem
  - problem: weights of actions may become too small
    $\rightarrow$ failure to explore
  - solution: occasionally, choose action uniformaly at random:

$$\hat{p}_i(t) = (1-\gamma)p_i(t) + \gamma/K$$

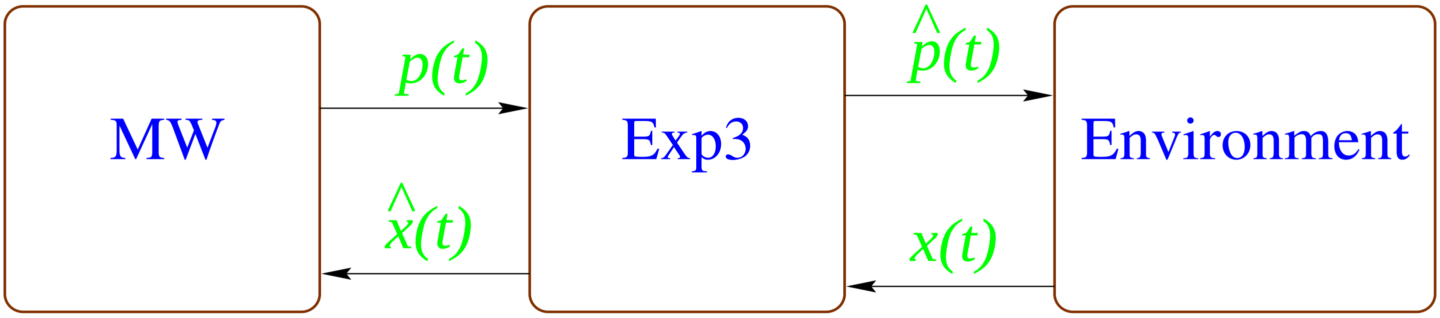
  - problem: only observe gain for one action
  - solution: simulate rewards for all actions

$$\hat{x}_i(t) = \begin{cases} x_{i_t}(t)/\hat{p}_{i_t}(t) & \text{if } i = i_t \\ 0 & \text{else} \end{cases}$$

    - $\mathbf{E}\left[\hat{x}_i(t) \mid i_1, \ldots, i_{t-1}\right] = x_i(t)$

An Algorithm for the Partial Information Game

- Exp3: "Exponential-weight algorithm for Exploration and Exploitation"



- parameters $\eta > 0$, $\gamma \in (0, 1)$
- for $t = 1, 2, \ldots T$
  - get distribution $\mathbf{p}(t)$ from MW
  - let $\hat{p}_i(t) = (1 - \gamma)p_i(t) + \gamma/K$
  - action i_t chosen randomly according to distribution $\hat{\mathbf{p}}(t)$
  - receive reward $x_{i_t}(t)$
  - let $\hat{x}_i(t) = \begin{cases} x_{i_t}(t)/\hat{p}_{i_t}(t) & \text{if } i = i_t \\ 0 & \text{else} \end{cases}$
  - feed rewards $\hat{x}_i(t)$ back to MW

Main Theorem

- Theorem: for appropriate parameter settings,

$$\mathbf{E}\left[G_{\text{Exp3}}\right] \geq \max_i \mathbf{E}\left[G_i\right] - 2.63\sqrt{TK\ln K}$$

- can replace T by any assumed upper bound on $\max_i \mathbf{E}\left[G_i\right]$
- so:

$$\mathbf{E}\left[\frac{G_{\text{Exp3}}}{T}\right] \geq \max_i \mathbf{E}\left[\frac{G_i}{T}\right] - 2.63\sqrt{\frac{K\ln K}{T}}$$

Example: Playing Repeated Games

- say playing unknown matrix $\mathbf{M} \in [0,1]^{n \times m}$ with value v
- on each round:
  - learner chooses distribution $\mathbf{p}(t)$ over rows
  - environment chooses column j_t
  - row i_t chosen according to $\mathbf{p}(t)$
  - learner receives payoff $\mathbf{M}(i_t, j_t)$
- can apply Exp3
  - $x_i(t) = \mathbf{M}(i, j_t)$
- Corollary: expected per-round payoff is at least:

$$v - 2.63\sqrt{\frac{n \ln n}{T}}$$

A Lower Bound

	regret
full information game	$O(\sqrt{T \ln K})$
partial information game	$O(\sqrt{TK \ln K})$

- can regret be improved?
  say, to
  $$O(T^\alpha (\ln K)^\beta)$$
  for some $\alpha < 1$, $\beta \geq 0$
- NO!
- Theorem: adversary can play so that every algorithm suffers expected regret
  $$\geq \Omega(\sqrt{KT})$$
- proof idea: adversary chooses rewards randomly using stationary distribution for each action
- stochastic case is worst case!

Bounding Regret with High Probability

- $\max_i \mathbf{E}[G_i]$ is weak measure of best-action payoff
- instead, compare actual payoff G_{Exp3} to actual best-action payoff $\max_i G_i$
- Theorem: for appropriate parameter settings, with probability $\geq 1 - \delta$

$$G_{\text{Exp3}} \geq \max_i G_i - 4.62 T^{2/3} (K \ln(K/\delta))^{1/3}$$

or

$$\frac{G_{\text{Exp3}}}{T} \geq \max_i \frac{G_i}{T} - 4.62 \left(\frac{K \ln(K/\delta)}{T} \right)^{1/3}$$

Using Expert Advice

- so far, only compare to best single action
- instead compare to more complicated strategies for choosing actions
  - e.g., all strategies which "switch" boundedly often between actions:

    "choose action 3 for rounds $1, \ldots, 37$
    action 17 for rounds $38, \ldots, 91$
    action 6 for rounds $92, \ldots, 200$"
- "expert" = agent advicing learner on actions to take
- learner wants to compete with best expert

Using Expert Advice (cont.)

- given access to N experts
- at time t
  - get action advised by each expert
  - combine into probability distribution $\mathbf{p}(t)$ over actions
- Exp4
  - "Exponential-weight algorithm for Exploration and Exploitation using Expert advice"
  - like Exp3, but maintain weights over experts (rather than actions)
- Theorem:

$$\mathbf{E}\left[G_{\text{Exp4}}\right] \geq \max_i \mathbf{E}\left[\text{total gain of expert } i\right] - 2.63\sqrt{TK \ln N}$$

- regret only logarithmic in N
- therefore, can have very large number of experts
  - e.g., all "switching" experts with at most r switches

Summary of Part II

- can use multiplicative-weights method for adversarial multi-armed bandit problem
  - Exp3 algorithm achieves expected regret $O(\sqrt{TK \ln K})$
  - slightly worse bound holds for actual regret
  - lower bound shows cannot be improved to $\text{poly}(\log K)$
  - but: can compete with best of N experts with only $O(\sqrt{TK \ln N})$ expected regret
- tractability of multi-armed bandit problem is not consequence of statistical assumptions
- can effectively learn to play games with very restricted feedback