Regret Minimization: Algorithms and Applications

Yishay Mansour
Google & Tel Aviv Univ.

Many thanks for my co-authors:
A. Blum, N. Cesa-Bianchi, and G. Stoltz

LEARNING

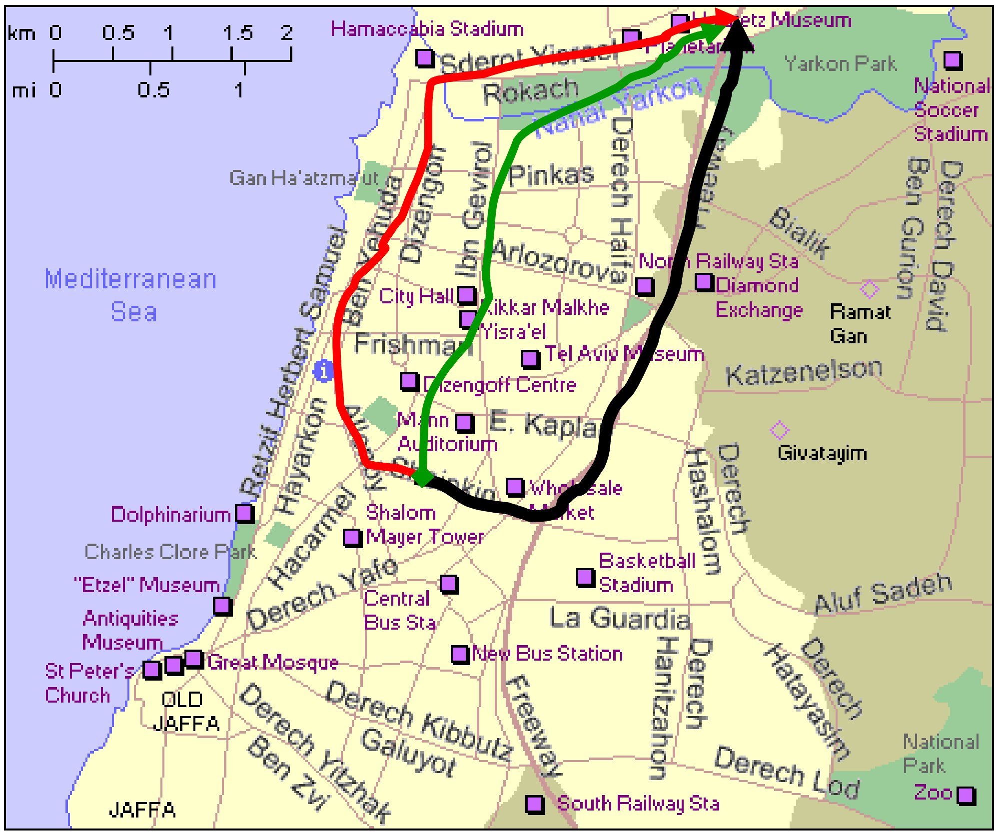
Weather Forecast

- Sunny:
- Rainy: 
- No meteorological understanding!
  - using other web sites

Web site	forecast
CNN	
BBC	
weather.com	
OUR	

Goal: Nearly the most accurate forecast

Route selection



Goal:

Fastest route

Challenge:

Partial Information

Rock-Paper-Scissors

			
	(0, 0)	(-1,1)	(1,-1)
	(1,-1)	(0, 0)	(-1,1)
	(-1,1)	(1,-1)	(0, 0)

Rock-Paper-Scissors

			
	(0, 0)	(-1,1)	(1,-1)
	(1,-1)	(0, 0)	(-1,1)
	(-1,1)	(1,-1)	(0, 0)

- Play multiple times
  - a repeated zero-sum game
- How should you “learn” to play the game?!
  - How can you know if you are doing “well”
  - Highly opponent dependent
  - In retrospect we should always win …

Rock-Paper-Scissors

			
	(0, 0)	(-1,1)	(1,-1)
	(1,-1)	(0, 0)	(-1,1)
	(-1,1)	(1,-1)	(0, 0)

- The (1-shot) zero-sum game has a value
  - Each player has a mixed strategy that can enforce the value
- Alternative 1: Compute the minimax strategy
  - Value V= 0
  - Strategy = (⅓, ⅓, ⅓)
- Drawback: payoff will always be the value V
  - Even if the opponent is "weak" (always plays )

Rock-Paper-Scissors

			
	(0, 0)	(-1,1)	(1,-1)
	(1,-1)	(0, 0)	(-1,1)
	(-1,1)	(1,-1)	(0, 0)

- Alternative 2: Model the opponent
  - Finite Automata
- Optimize our play given the opponent model.
- Drawbacks:
  - What is the “right” opponent model.
  - What happens if the assumption is wrong.

Rock-Paper-Scissors

			
	(0, 0)	(-1,1)	(1,-1)
	(1,-1)	(0, 0)	(-1,1)
	(-1,1)	(1,-1)	(0, 0)

- Alternative 3: Online setting
  - Adjust to the opponent play
  - No need to know the entire game in advance
  - Payoff can be more than the game's value V
- Conceptually:
  - Have a set of comparison class of strategies.
  - Compare performance in hindsight

Rock-Paper-Scissors

(0, 0)	(-1,1)	(1,-1)
(1,-1)	(0, 0)	(-1,1)
(-1,1)	(1,-1)	(0, 0)

- Comparison Class H:
  - Example : A={  }
  - Other plausible strategies:
    - Play what you opponent played last time
    - Play what will beat your opponent previous play
- Goal:

Online payoff near the best strategy in the class H

- Tradeoff:
  - The larger the class H, the difference grows.

Rock-Paper-Scissors: Regret

- Consider A={  ,  ,  }
  - All the pure strategies
- Zero-sum game:

  Given any mixed strategy σ of the opponent,

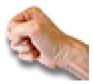
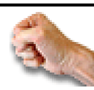
  there exists a pure strategy $a \in A$

  whose expected payoff is at least V
- Corollary:

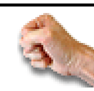
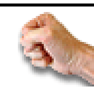
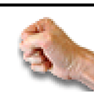
  For any sequence of actions (of the opponent)

  We have some action whose average value is V

Rock-Paper-Scissors: Regret

we	opponent	payoff
		-1
		1
		0
		0
		-1
		-1

Average payoff -1/3

play 	opponent	payoff
		1
		0
		0
		1
		1
		-1

New average payoff 1/3

Rock-Paper-Scissors: Regret

- More formally:

  After T games,

  $\hat{U}$ = our average payoff,

  U(h) = the payoff if we play using h

  regret(h) = U(h)- $\hat{U}$

- Claim:

If for every $a \in A$ we have $regret(a) \leq \varepsilon$, then $\hat{U} \geq V - \varepsilon$

- External regret: $\mathbf{max}_{h \in H}\, regret(h)$

REGRET

MINIMIZATION

[Blum & M] and [Cesa-Bianchi, M & Stoltz]

Regret Minimization: Setting

- Online decision making problem (single agent)
- At each time, the agent:
  - selects an action
  - observes the loss/gain
- Goal: minimize loss (or maximize gain)
- Environment model:
  - *stochastic* versus *adversarial*
- Performance measure:
  - *optimality* versus *regret*

Regret Minimization: Model

- Actions $A=\{1, \ldots ,N\}$
- Number time steps: $t \in \{ 1, \ldots , T\}$
- At time step t:
  - The agent selects a distribution p_i^t over A
  - Environment returns costs $c_i^t \in [0,1]$
  - Online loss: $l^t = \Sigma_i c_i^t \; p_i^{\mathrm{t}}$
- Cumulative loss : $L_{online} = \Sigma_t l^t$
- Information Models:
  - Full information: observes every action's cost
  - Partial information: observes only it own cost

Stochastic Environment

- Costs: c_i^t are *i.i.d.* random variables
  - Assuming an oblivious opponent
- Tradeoff: Exploration versus Exploitation
- Approximate solution:
  - sample each action *O(logT)* times
  - select the best observed action
- Gittin's Index
  - Simple optimal selection rule
    - under some Bayesian assumptions

Competitive Analysis

- Costs: c_i^t are generated adversarially,
  - might depend on the online algorithm decisions
    - in line with our game theory applications
- Online Competitive Analysis:
  - Strategy class = any dynamic policy
  - too permissive
    - Always wins rock-paper-scissors
- Performance measure:
  - compare to the best strategy in a class of strategies

External Regret

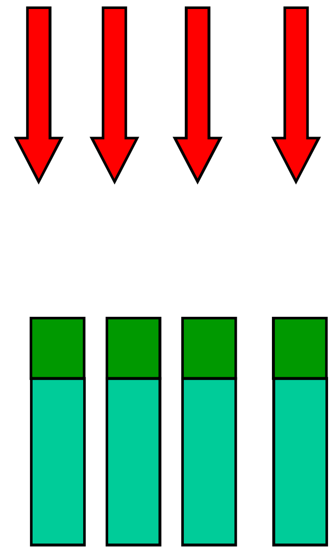
- Static class
  - Best fixed solution
    - Compares to a single best strategy (in *H*)
- The class *H* is fixed beforehand.
  - optimization is done with respect to *H*
- Assume $H=A$
  - Best action: $L_{best} = MIN_i \{\Sigma_t\, c_i^t\}$
  - External Regret = $L_{online} - L_{best}$
    - Normalized regret is divided by T

External regret: Bounds

- Average external regret goes to zero
  - No regret
  - Hannan [1957]
- Explicit bounds
  - Littstone & Warmuth '94
  - CFHHSW '97
  - External regret = $O(\sqrt{T \log N})$

External Regret: Greedy

- Simple Greedy:
  - Go with the best action so far.
- For simplicity loss is *{0,1}*
- Loss can be N times the best action
  - holds for any deterministic online algorithm

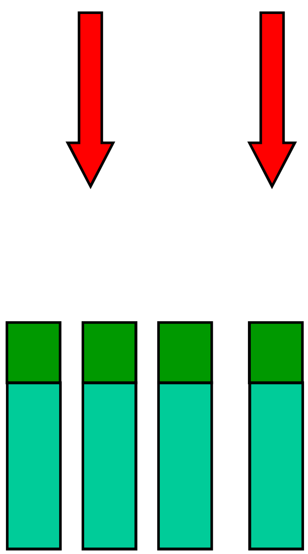


External Regret: Randomized Greedy

- Randomized Greedy:
  - Go with a *random* best action.
- Loss is *ln(N)* times the best action
- Analysis:

  When the ***best*** increases from k to $k+1$ expected loss is

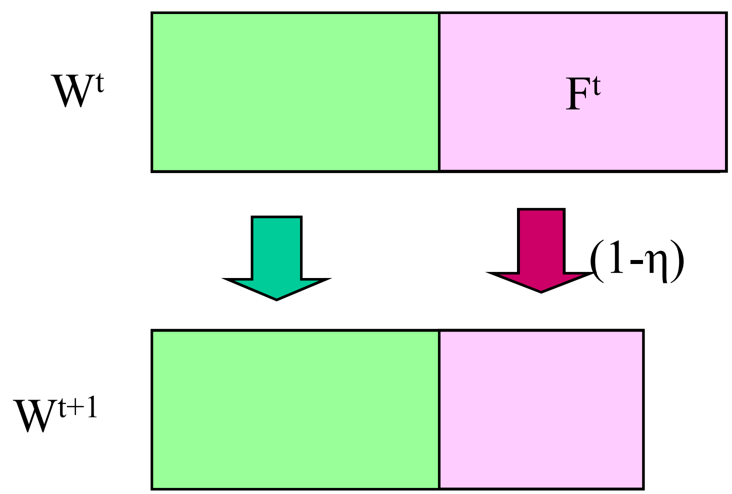
  $1/N + 1/(N-1) + \ldots \approx ln(N)$



External Regret: PROD Algorithm

- Regret is $\sqrt{T \log N}$
- PROD Algorithm:
  - plays sub-best actions
  - Uses exponential weights

    $w_a = (1\text{-}\eta)^{L_a}$
  - Normalize weights
- Analysis:
  - W^t= weights of all actions at time t
  - F^t= fraction of weight of actions with loss **1** at time t
    - Also, expected loss: $L_{ON} = \sum F_t$



$$\boxed{W^{t+1} = W^t\,(1\text{-}\eta F^t)}$$

External Regret: Bounds Derivation

- Bounding W^T
- Lower bound:

$$W^T > (1-\eta)^{L_{min}}$$

- Upper bound:

$$\begin{aligned} W^T &= W^1 \Pi_t (1-\eta F^t) \\ &\leq W^1 \Pi_t \exp\{-\eta F^t\} \\ &= W^1 \exp\{-\eta L_{ON}\} \end{aligned}$$

using $1-x \leq e^{-x}$

- Combined bound:

$$(1-\eta)^{L_{min}} \leq W^1 \exp\{-\eta L_{ON}\}$$

- Taking logarithms:

$$L_{min} \log(1-\eta) \leq \log(W^1) - \eta L_{ON}$$

- Final bound:

$$L_{ON} \leq L_{min} + \eta L_{min} + \log(N)/\eta$$

- Optimizing the bound:

$$\eta = \sqrt{\log(N)/L_{min}}$$

$$L_{ON} \leq L_{min} + 2\sqrt{L_{min} \log(N)}$$

External Regret: Summary

- We showed a bound of $2\sqrt{L_{min} \log N}$
- More refined bounds

  $\sqrt{Q \log N}$ where $Q = \Sigma_t (c^t_{best})^2$
- More elaborate notions of regret …

External Regret: Summary

- How surprising are the results …
  - Near optimal result in online adversarial setting
    - very rear …
  - Lower bound: stochastic model
    - stochastic assumption do not help …
  - Models an “improved” greedy
  - An “automatic” optimization methodology
    - Find the best fixed setting of parameters

Internal Regret

Internal Regret

- Game theory applications:
  - Avoiding dominated actions
  - Correlated equilibrium
- Reduction from External Regret [Blum & M]

Dominated actions

- Action a_i is dominated by b_i if for every a^{-i} we have $u_i(a_i, a^{-i}) < u_i(b_i, a^{-i})$
- Clearly, we like to avoid dominated action
  - Remark: an action can be also dominated by a mixed action
- Q: can we guarantee to avoid dominated actions?!

1	2	0	3	1	a
2	5	1	9	12	b

Dominated Actions & Internal Regret

- How can we test it?!
  - in retrospect
- a_i is dominates b_i
  - Every time we played a_i we do better with b_i
- Define internal regret
  - swapping a pair of actions
- No internal regret → no dominated actions

our actions	Our Payoff	Modified Payoff (a→b)	Internal Regret (a→b)
a	1	2	2-1=1
b			
c			
a	2	5	5-2=3
d			
a	3	9	9-3=6
b			
d			
a	0	1	1-0=1

Dominated actions & swap regret

- Swap regret
  - An action sequence $\sigma = \sigma_1, \ldots, \sigma_t$
  - Modification function $F: A \rightarrow A$
  - A modified sequence

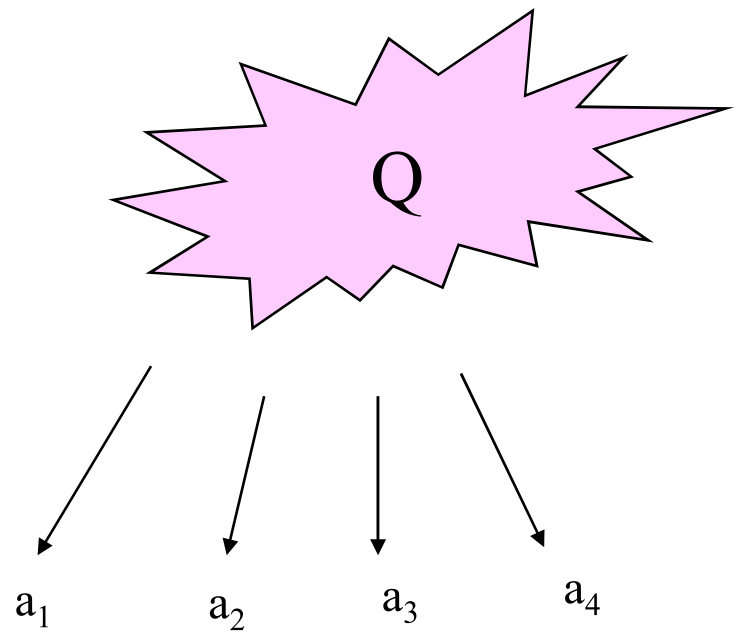
    $\sigma(F) = F(\sigma_1), \ldots, F(\sigma_t)$
- Swap_regret =

  $\max_F V(\sigma(F)) - V(\sigma)$
- Theorem: If Swap_regret < R then in at most R/ε steps we play ε-dominated actions.

σ	σ(F)
a	b
b	c
c	c
a	b
d	b
a	b
b	c
d	b
a	b

Correlated Equilibrium

- Q a distribution over joint actions
- Scenario:
  - Draw joint action *a* from Q,
  - player *i* receives action a_i
    - and no other information
- Q is a correlated Eq if:
  - for every player *i*, the recommended action a_i is a best response
    - given the induced distribution.



Swap Regret & Correlated Eq.

- Correlated Eq ⇔ NO Swap regret
- Repeated game setting
- Assume swap_regret $\leq \varepsilon$
  - Consider the empirical distribution
    - A distribution Q over joint actions
  - For every player it is ε best response
  - Empirical history is an ε correlated Eq.

Internal/Swap Regret

- Comparison is based on online's decisions.
  - depends on the actions of the online algorithm
  - modify a single decision (consistently)
    - Each time action A was done do action B
- Comparison class is not well define in advanced.
- Scope:
  - Stronger then External Regret
  - Weaker then competitive analysis.

Internal & Swap Regret

- Assume action sequence $A=a_1 \ldots a_T$
  - Modified input $(b \rightarrow d)$:
    - Change every $a_i^t=b$ to $a_i^t=d$, and create actions seq. B.
    - $L(b\rightarrow d)$ is the cost of B
      - using the same costs c_i^t
- Internal regret

$$L_{\text{online}}-\min_{\{b,d\}} L_{online}(b\rightarrow d) = \max_{\{b,d\}} \Sigma_t (c_b^t - c_d^t) p_b^t$$

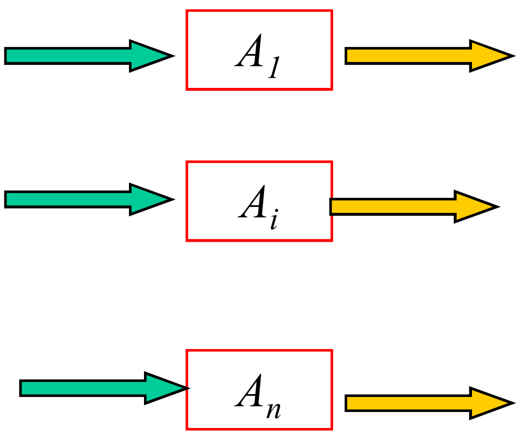
- Swap Regret:
  - Change action i to action $F(i)$

Internal regret

- No regret bounds
  - Foster & Vohra
  - Hart & Mas-Colell
  - Based on the approachability theorem
    - Blackwell ’56
  - Cesa-Bianchi & Lugasi ’03
    - Internal regret = $O(\log N + \sqrt{T \log N})$

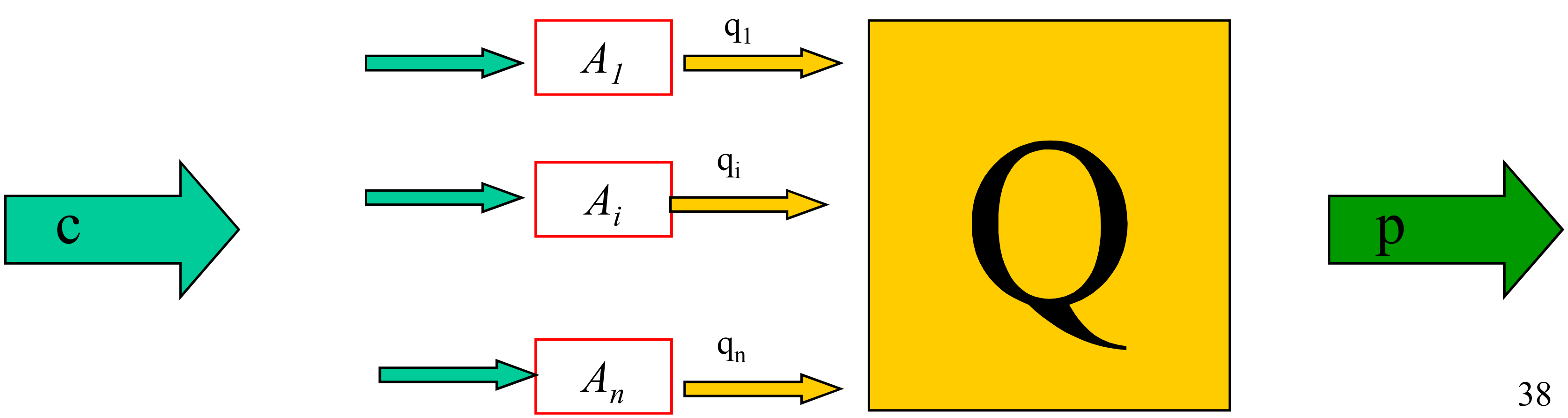
External Regret to Internal Regret: Generic reduction

- Input:
  - N (External Regret) algorithms
  - Algorithm A_i, for any input sequence :
    - $L_{Ai} \leq L_{best,i} + R_i$

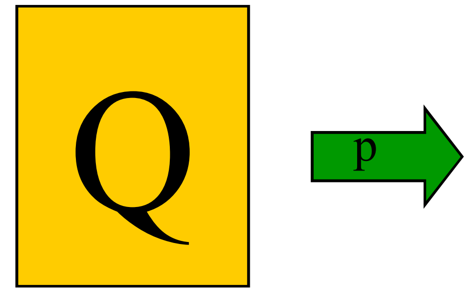


External to Internal: Generic reduction

- General setting (at time t):
  - Each A_i outputs a distribution q_i
    - A matrix Q
  - We decide on a distribution p
  - Adversary decides on costs $c=<c_1 \ldots c_N>$
  - We return to A_i some cost vector



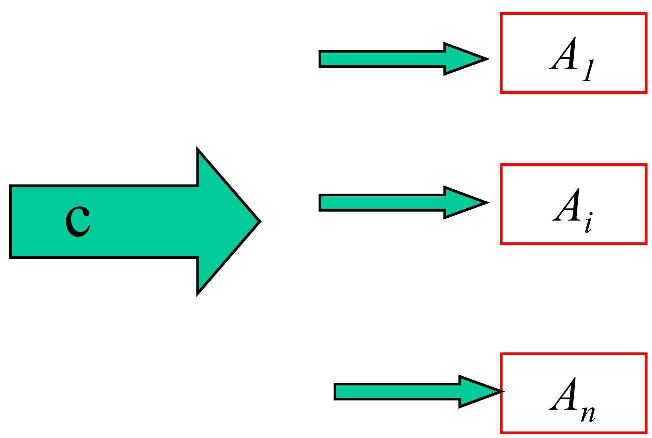
Combining the experts



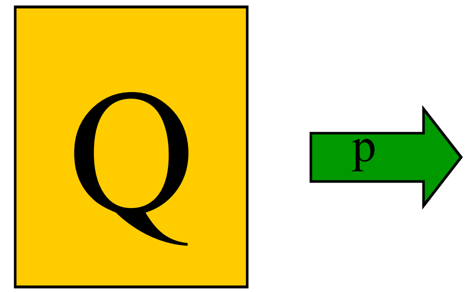
- Approach I:
  - Select an expert A_i with probability r_i
  - Let the "selected" expert decide the outcome
  - action distribution $p=Qr$
- Approach II:
  - Directly decide on p.
- Our approach: make $p=r$
  - Find a p such that $p=Qp$

Distributing loss

- Adversary selects costs $c=<c_1 \ldots c_N>$
- Reduction:
  - Return to A_i cost vector $c_i = p_i\, c$
  - Note: $\Sigma\, c_i = c$



External to Internal: Generic reduction



- Combination rule:
  - Each A_i outputs a distribution q_i
    - Defines a matrix Q
  - Compute p such that $p=Qp$
    - $p_j = \Sigma_i \, p_i \, q_{i,j}$
  - Adversary selects costs $c=<c_1 \ldots c_N>$
  - Return to A_i cost vector $p_i \, c$

Motivation

- Dual view of P:
  - p_i is the probability of selecting action i
  - p_i is the probability of selecting algo. A_i
    - Then use A_i probability, namely q_i
- Breaking symmetry:
  - The feedback to A_i depends on p_i

Proof of reduction:

- Loss of A_i (from its view)
  - $<(p_i\, c), q_i> = p_i < q_i\, , c>$
- Regret guarantee (for any action i):
  - $L_i = \Sigma_t\, p_i^t <q_i^t, c^t> \;\leq\; \Sigma_t\, p_i^t\, c_j^t + R_i$
- Online loss:
  - $L_{online} \quad = \Sigma_t <p^t, c^t>$

    $= \Sigma_t <p^t\, Q^t, c^t>$

    $= \Sigma_t\, \Sigma_i\, p_i^t <q_i^t, c^t> = \Sigma_i\, L_i$
- For any swap function F:
  - $L_{online} \leq L_{online,F} + \Sigma_i\, R_i$

Swap regret

- Corollary: For any swap F:

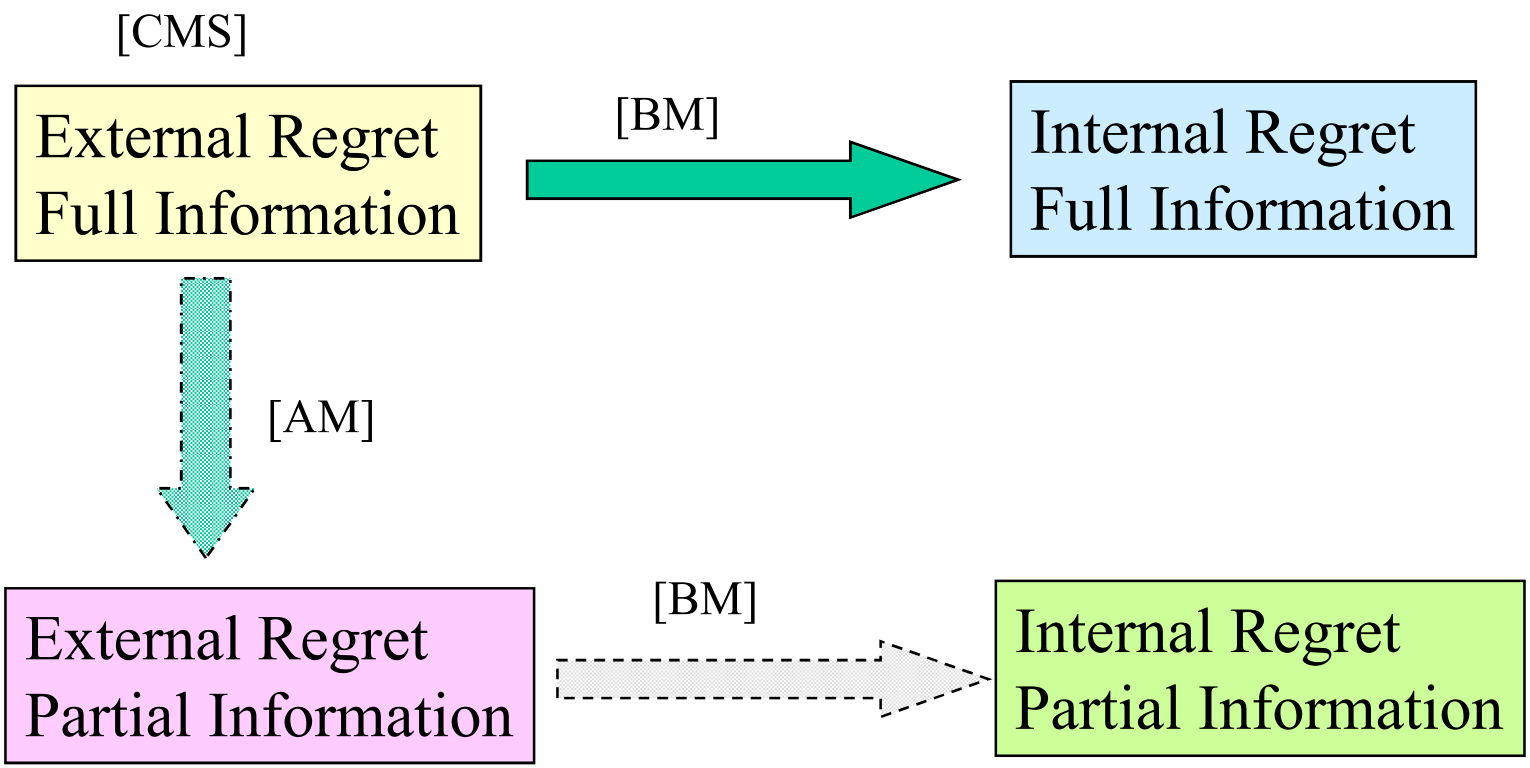
$$L_{online} \le L_{online,F} + O(N\sqrt{T\log(N)} + N\log(N))$$

- Improved bound:
  - Note that $\Sigma_i L_{max,i} \le T$
    - worse case all $L_{max,i}$ are equal.
  - Improved bound:

$$L_{online} \le L_{online,F} + O\left(\sqrt{TN\log(N)}\right)$$

Summary

Reductions between Regrets



More elaborate regret notions

- Time selection functions [Blum & M]
  - determines the relevance of the next time step
  - identical for all actions
  - multiple time-selection functions
- Wide range regret [Lehrer, Blum & M]
  - Any set of modification functions
    - mapping histories to actions

Conclusion and Open problems

- Reductions
  - External to Internal Regret
    - full information
    - partial information
- SWAP regret Lower Bound
  - poly in N= |A|
    - Very weak lower bounds
- Wide Range Regret
  - Applications …

Thank You!

Many thanks for my co-authors:
A. Blum, N. Cesa-Bianchi, and G. Stoltz