

# Sequential decisions under uncertainty

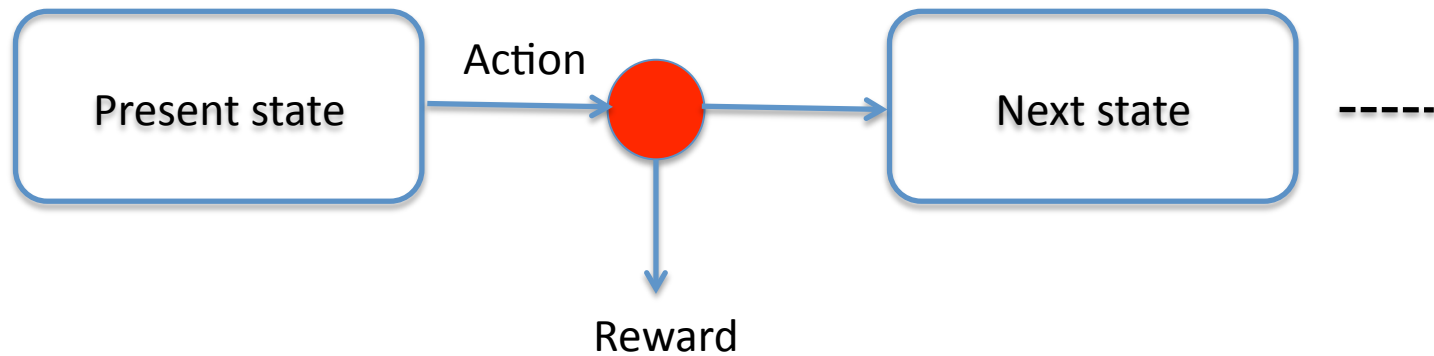
KTH/EES PhD course

Lecture 7

# Lecture 7

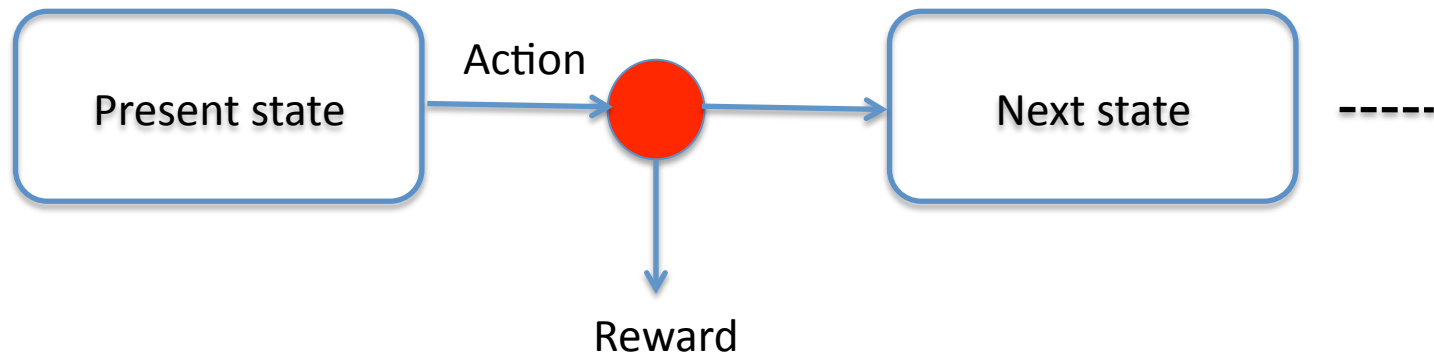
- MDP: extensions
  - POMDP
  - Limit theorems
- Multi-armed bandit problems
  - Introduction
  - Lower regret bound

# POMDP



- Set of states:  $S$
- Set of actions available in state  $s$ :  $A_s$ ,  $A = \cup_{s \in S} A_s$
- These sets are finite, countably infinite, or compact subsets of a Euclidian space (finite dimension)
- Time horizon

# Observations



- The state is not fully observable
- Observation at time  $t$ :  $O_t$
- Observation probabilities:

$$o(a, s, z) = P[O_t = z | X_t = s, Y_{t-1} = a]$$

# Decision rules, policies

- HR:  $\pi = (\pi_1, \dots, \pi_{N-1})$   
 $\pi_t : (\mathcal{Z} \times \mathcal{A})^{t-1} \times \mathcal{Z} \rightarrow \mathcal{P}(\mathcal{A})$   
 $q_{\pi_t(h_t)}(a) : \text{probability to select action } a$
- Main issue: Markovian policies have poor performance
- Optimal policies are history-based!

# Information states

- An information state is a distribution on the state space: it is our belief for the state distribution, given the history
- Bayesian approach. Updating belief distribution:
  - At time  $t$ : belief  $b$
  - At time  $t+1$ , given that action  $a$  was chosen, and observation  $z$  has just been made,

$$b_z^a(s') = \frac{o(a, s', z) \sum_s p(s'|s, a)b(s)}{\sum_{s, s''} o(a, s'', z)p(s''|s, a)b(s)}$$

- Markovian structure recovered (MDP)

# More on POMDP

- Read Anthony Cassandra's thesis:  
"Exact and Approximate Algorithms for POMDP"

# Limit theorems

- Read Kushner Dupuis book:  
“Numerical methods for stochastic control problems in continuous time”
- ... or Kushner’s paper (same title), SIAM J. Control and Optimization, 1990

# An example

- Finite time horizon (N steps), and finite “budget” B

$$b(t + 1) = b(t) - c(a(t), \xi(t))$$

$$b(0) = B$$

- Reward:  $r(a(t), \xi(t))$

- Goal: max expected reward

- Value function:  $v(b, t) = \sup_u \mathbb{E} \left[ \sum_{i=0}^{t-1} r(u(i), \xi(i)) \right]$

- Scaling budget and time-horizon:  $V_\beta(B, T) = \beta v(B/\beta, t/\beta)$

$$\beta \rightarrow 0$$

# An example

- Result:  $V(B, T) = \lim_{\beta \rightarrow 0} V_\beta(B, T)$  exists and solves

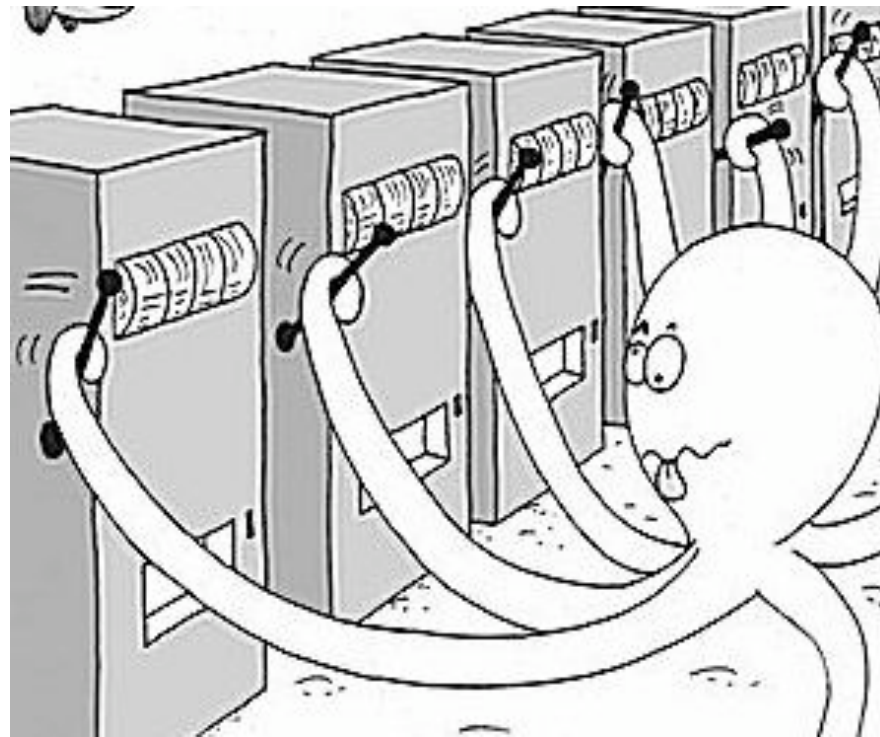
$$V(B, T) = \sup_u \int_0^{\tau \wedge T} \bar{r}(u(t)) dt$$

$$\tau = \inf\{t \geq 0 : B(t) = 0\}$$

$$\frac{dB}{dT} = -\bar{c}(u(t))$$

$$\bar{r}(a) = \mathbb{E}_\xi[r(u, \xi(0))], \quad \bar{c}(a) = \mathbb{E}_\xi[c(u, \xi(0))]$$

# Multi-Armed Bandit (MAB)



# MAB problem

- Known parameters: number  $K$  of arms (or decisions), time horizon (or number of rounds)  $T$
- Unknown parameters: how rewards are generated  
 $X_{j,t}$  : reward of pulling arm  $j$  at time  $t$
- Objective: maximize the total expected reward at time  $T$

# Stochastic vs. Adversarial

- Stochastic: rewards sampled from an unknown distribution
  - Example: IID case,  
 $(X_{j,t}, t = 1, 2, \dots)$  IID random variables with mean  $\mu_j$
- Adversarial setting: rewards chosen by an adversary
  - Oblivious adversary:  
 $(X_{j,t}, t = 1, 2, \dots)$  chosen initially (at time 0)
  - Adaptive adversary: rewards depend on the history (selected arms so far)

# Applications

- Clinical trials (Thompson 1933)
- Ads placement on webpages
- Routing problems
- ...

# Outline of the next lectures

- Asymptotically optimal policies for IID MAB + UCB policies
- Finite-time analysis of IID MAB
- Large number of arms
  - Unstructured rewards
  - Structured rewards
- Adversarial MAB

# Stochastic MAB

- Robbins 1952
- IID rewards

$(X_{j,t}, t = 1, 2, \dots)$  IID random variables with mean  $\mu_j$

- At a given time, an arm is selected and the corresponding random reward is observed
- Best arm:  $j^* = \arg \max_j \mu_j$
- Under a given policy, the arm selected at time  $t$  is  $j(t)$

Expected regret:

$$R(t) = t \times \mu_{j^*} - \sum_{n=1}^t \mu_{j(n)}$$

# Parametric model

- Measure on  $\mathbb{R}$  :  $\nu$
- Reward distributions parametrized by  $\theta \in \mathbb{R}$
- Configuration:  $C = (\theta_1, \dots, \theta_K)$
- Arm  $j$  reward distribution:  $X_{j,t} \sim f(x, \theta_j)d\nu(x)$

$$\int |x|f(x, \theta_j)d\nu(x) < \infty$$

$$\int xf(x, \theta_j)d\nu(x) = \mu(\theta_j)$$

- Kullback-Leibler divergence:

$$I(\theta, \lambda) = \int \log \left[ \frac{f(x, \theta)}{f(x, \lambda)} \right] f(x, \theta)d\nu(x)$$

# Assumptions

- $\mu(\theta)$  strictly increasing
- $I(\theta, \lambda)$  continuous in  $\lambda$

$$\theta \neq \lambda \implies I(\theta, \lambda) > 0$$

- Finally:  $\forall \lambda, \forall \delta, \exists \lambda' :$

$$\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta$$

- Notation: permutation  $\sigma$

$$\mu(\theta_{\sigma(1)}) \geq \dots \geq \mu(\theta_{\sigma(K)})$$

$$\mu(\theta_{\sigma(1)}) = \mu(\theta_{\sigma(l)}) > \mu(\theta_{\sigma(l+1)})$$

# Example: Bernoulli rewards

- Rewards take values in  $\{0,1\}$
- Measure  $\nu$ :  $\nu = \delta_0 + \delta_1$
- We have:  $\theta \in [0, 1]$

$$\mu(\theta) = \theta$$

$$I(\theta, \lambda) = \theta \log \left[ \frac{\theta}{\lambda} \right] + (1 - \theta) \log \left[ \frac{1 - \theta}{1 - \lambda} \right]$$

# Regret and uniformly good rules

- Number of time arm  $j$  selected up to time  $t$ :  $T_t(j)$
- Expected regret:

$$R(t, C) = \sum_{j \notin \{\sigma(1), \dots, \sigma(l)\}} (\mu(\theta_{\sigma(1)}) - \mu(\theta_j)) \mathbb{E}[T_t(j)]$$

- Uniformly good rule: for all configuration  $C$

$$\mathbb{E}[T_t(j)] = o(t^\alpha), \quad \forall \alpha > 0, \forall j \notin \{\sigma(1), \dots, \sigma(l)\}$$

# Lower bound on regret

Lai and Robbins 1985

**Theorem** Consider any uniformly good rule.

Configuration:  $C = (\theta_1, \dots, \theta_K)$

$\forall \epsilon > 0, \quad \forall j \notin \{\sigma(1), \dots, \sigma(l)\},$

$$\lim_{t \rightarrow \infty} P_C \left[ T_t(j) \geq \frac{(1 - \epsilon) \log t}{I(\theta_j, \theta_{\sigma(1)})} \right] = 1.$$

Hence:

$$\liminf_{t \rightarrow \infty} \frac{R(t, C)}{\log(t)} \geq \sum_{j \notin \{\sigma(1), \dots, \sigma(l)\}} \frac{\mu(\theta_{\sigma(1)}) - \mu(\theta_j)}{I(\theta_j, \theta_{\sigma(1)})}.$$

# Universality of the bound

- Similar bound can be derived for controlled Markov chains, i.e., for parametrized average reward MDP
- Graves-Lai 1996. Asymptotically efficient adaptive choice of control laws in controlled Markov chains.

# Model

- Markov chain:  $X_n, n \geq 0$
- Action space  $A$
- Transition probabilities:  $p(y|x, a, \theta)$
- Unknown parameter:  $\theta$
- Stationary control laws:  $G = (g_1, \dots, g_K)$
- Under control law  $g$ , irreducible MC, with stationary distribution  $\pi_\theta^g$
- Reward:  $\mu_\theta(g) = \int r(x, g(x)) d\pi_\theta^g(x)$

$$\mu^* = \max_g \mu_\theta(g)$$

# Lower bound on regret

- The regret can be shown to “look” like:

$$R(t, \theta) = \sum_{g: \mu_\theta(g) < \mu^*} (\mu^* - \mu_\theta(g)) \mathbb{E}[T_t(g)]$$

- We have:  $\liminf_{t \rightarrow \infty} R(t, \theta) \geq c(\theta)$

$$c(\theta) = \inf \left\{ \frac{\sum_{j \notin J(\theta)} \alpha_j (\mu^* - \mu_\theta(g_j))}{\inf_{\lambda \in B(\theta)} \sum_{j \notin J(\theta)} \alpha_j I^{g_j}(\theta, \lambda)} : \sum_{j \notin J(\theta)} \alpha_j = 1 \right\}$$

$J(\theta)$  : set of optimal control laws for parameter  $\theta$

$B(\theta)$  : set of parameters such the optimal control laws under  $\theta$  are not optimal, and cannot be “distinguished”