Markovian Decision Process (MDP):
theory and applications to wireless networks

Philippe Ciblat

Joint work with I. Fawaz, N. Ksairi, C. Le Martret, M. Sarkiss
Outline

- A few examples

- Markov Decision Process
  - Markov chain
  - Discounted approach
  - Bellman's equation and fixed-point theorem
  - Practical algorithms

- Applications to wireless networks
  - Channel exploration
  - Hybrid ARQ optimization
  - Scheduling with Energy Harvesting
Part 1 : A few examples
Example 1: Inventory control

- $x_k$ stock at the beginning of slot $k$
- $a_k$ stock ordered and immediately delivered at the beginning of slot $k$
- $w_k$ request of stock during slot $k$ (iid process)

Goal: optimizing $\{a_k\}$? but

$a_k$ too large (cost); $a_k$ too small (stock outage with $x_{k+1} = 0$)
Example 1: Inventory control (cont’d)

Mathematical model 1:
- Reward: \( r(x_k, a_k) \)

\( \alpha \)-Discounted (long-term) reward: 
\[
 r = \lim_{N \to \infty} \sum_{n=0}^{N} \alpha^n r(x_n, a_n)
\]

Additive cost over time
- Sequential dynamic:
\[
 x_{k+1} = f(x_k, a_k, w_k) = (x_k + a_k - w_k)^+
\]

\( \{x_k\} \) Markov chain with transition probability \( Q(x_{k+1} | x_k, a_k) \)
- Policy: \( a_{k+1} = \mu(k)(x_k) \)

\[
 \{\mu(k)\}? \min_{\mu} \mathbb{E}_{|x_0, \mu} \left[ \lim_{N \to \infty} \sum_{n=0}^{N} \alpha^n r(x_n, \mu(n)(x_n)) \right]
\]

Markov Decision Process (MDP)
Example 1: Inventory control (cont’d)

Mathematical model 2

- Reward/Cost: $c(x_k)$
- Outage: $\tilde{x}_k = 0$ if $x_k + a_k - w_k \geq 0$, $1$ otherwise

$$o(\tilde{x}_k) = \tilde{x}_k \Rightarrow \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} o(\tilde{x}_n) < O_t$$

- Let $s_k = [x_k, \tilde{x}_k]$: system state. Optimal policy:

$$\{\mu(k)\} \in \underset{\mu}{\text{arg min}} \mathbb{E}_{s_0, \mu} \left[ \lim_{N\to\infty} \sum_{n=0}^{N} \alpha^n r(s_n, \mu(n)(s_n)) \right]$$

s.t.

$$\mathbb{E}_{s_0, \mu} \left[ \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} o(s_n) \right] < O_t$$

Constrained Markov Decision Process (CMDP)
Example 2: Water reservoir control

Reservoir with \( C \) finite capacity: maximize simultaneously the stock and water release

- \( x_k \) water volume at time \( k \)
- \( a_k \) water released volume at time \( k \) (for electricity or irrigation)
- \( w_k \) random inflow during slot \( k \) (rain and tributary rivers): iid process

We have

\[
x_{k+1} = \min\left(x_k - a_k + w_k, C\right)
\]

- Cost: \( r(x_k, a_k) = x_k + a_k \)
- \( x_k \) not perfectly known, but estimated. So we know

\[
y_k = g(x_k, \zeta_k)
\]

**Partially Observable Markov Decision Process (POMDP)**
Part 2 : Markov Decision Process
Let \( \{x_k\} \) a random sequence/process.

**Definition**

A sequence is said a Markov chain iff

\[
P(x_{k+1} | x_k, \ldots, x_0) = P(x_{k+1} | x_k), \ \forall k
\]

If \( x_k \) takes a finite number of values \( \mathcal{X} = \{x^{(1)}, \ldots, x^{(V)}\} \),

- transition probability matrix:
  \[
  T = (t_{m,n})_{1 \leq m, n \leq V}
  \text{ with }
  t_{m,n} = \text{Prob}(x_{k+1} = x^{(m)} | x_k = x^{(n)}) \geq 0
  \text{ and } \sum_{m=1}^{V} t_{m,n} = 1.
  \]
- Graph theory
- (Stochastic) non-negative matrix theory
State, Action, and Policy

- **State**: $x_k$ (system information at time $k$)
- **Action**: $a_k$
- **Disturbance**: $w_k$ iid process

\[ x_{k+1} = f(x_k, a_k, w_k) \]

**Transition kernel**:

\[ Q(x_{k+1} | x_k, a_k) = \text{Prob}(x_{k+1} = x^{(m)} | x_k = x^{(n)}, a_k = a) = \int t_{m,n}(a, w)p_w(w)dw \]

- **Reward**: $r(x_k, a_k)$

\[ \mathbb{E}_{x_0} \left[ \sum_{n=0}^{\infty} \alpha^n r(x_n, a_n) \right] \quad (\text{discount}) \quad \text{or} \quad \mathbb{E}_{x_0} \left[ \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N} r(x_n, a_n) \right] \]

- **Policy**: $\mu_k(x_k) \in \mathcal{A}$ with $\mathcal{A}$ set of actions
  - deterministic ($\mu_k$ is a function) or random ($\mu_k$ is a pdf)
  - stationary ($\mu_k$ independent of $k$) or nonstationary
Bellman’s equation

- Let $\mu$ be a deterministic stationary policy
- Let $R_\mu(x_0)$ be the discounted reward assuming used policy $\mu$ and initialized at $x_0$

\[
R_\mu(x_0) = \sum_{n=0}^{\infty} \alpha^n r(x_n, \mu(x_n))
\]

\[
= r(x_0, \mu(x_0)) + \alpha \sum_{n=0}^{\infty} \alpha^n r(x_{n+1}, \mu(x_{n+1}))
\]

\[
= r(x_0, \mu(x_0)) + \alpha \sum_{n=0}^{\infty} \alpha^n r(x_{n+1}, \mu(x_{n+1}))
\]

\[
= r(x_0, \mu(x_0)) + \alpha \int R_\mu(x) Q(x|x_0, \mu(x_0)) dx, \quad \forall x_0
\]

Bellman’s equation

\[
R_\mu = T_\mu(R_\mu) \quad \text{(fixed-point)}
\]

\[
x_0 \mapsto T_\mu(f)(x_0) = r(x_0, \mu(x_0)) + \alpha \int f(x) Q(x|x_0, \mu(x_0)) dx
\]
We also define

\[ T(f)(x_0) = \max_{a \in \mathcal{A}} \left\{ r(x_0, a) + \alpha \int f(x)Q(x|x_0, a)dx \right\} \]

**Lemma**

It exits an unique function \( R \) (resp. \( R_\mu \)) such that

\[ R = T(R) \quad \text{and} \quad R_\mu = T_\mu(R_\mu) \]

**Sketch of proof :**

- Let \( \|f\|_\infty = \sup_x |f(x)| \) (within set of bounded function)
- Assuming \( r(\cdot, \cdot) \) bounded function for any state and action

\[
\| T(f) - T(g) \|_\infty \leq \alpha \max_{a \in \mathcal{A}} \left\{ \int f(x)Q(x|x_0, a)dx - \int g(x)Q(x|x_0, a)dx \right\} \\
\leq \alpha \|f - g\|_\infty
\]

\( T \) is a \( \alpha \)-contraction, so Banach’s fixed-point theorem applies
A (deterministic stationary) policy $\mu^*$ is said optimal

$R_{\mu^*}(x_0) = R^*(x_0), \ \forall x_0 \text{ with } R^*(x_0) = \sup_\mu R_\mu(x_0)$

Equivalently, $R_{\mu^*}$ is the fixed point of $T$

**Sketch of proof**: Let $\mu^*$ a policy s.t. $R_{\mu^*}$ is the fixed point of $T$

- $R_{\mu^*}(x_0) \geq r(x_0, a) + \alpha \int R_{\mu^*}(x)Q(x|x_0, a)dx, \ \forall (x_0, a) \in X \times A$
- $R_{\mu^*}(x_n) \geq r(x_n, a_n) + \alpha \int R_{\mu^*}(x)Q(x|x_n, a_n)dx$

$\mathbb{E}_{x_0, \mu}[\alpha^n R_{\mu^*}(x_n)] - \alpha^{n+1} \mathbb{E}_{x_0, \mu}[\int R_{\mu^*}(x)Q(x|x_n, a_n)dx] \geq \mathbb{E}_{x_0, \mu}[\alpha^n r(x_n, a_n)] \times \alpha^n + \text{expectation}$

$\mathbb{E}_{x_0, \mu}[\alpha^n \int R_{\mu^*}(x)Q(x|h_{n-1})dx] - \mathbb{E}_{x_0, \mu}[\alpha^{n+1} \int R_{\mu^*}(x)Q(x|h_n)dx] \geq \mathbb{E}_{x_0, \mu}[\alpha^n r(x_n, a_n)]$ (Markov)

$R_{\mu^*}(x_0) - \alpha^{N+1} \mathbb{E}_{x_0, \mu}[\int R_{\mu^*}(x)Q(x|h_N)dx] \geq \mathbb{E}_{x_0, \mu}[\sum_{n=0}^{N} \alpha^n r(x_n, a_n)]$ (sum over $n$)

$R_{\mu^*}(x_0) \geq \mathbb{E}_{x_0, \mu}[\sum_{n=0}^{\infty} \alpha^n r(x_n, a_n)] = R_\mu(x_0), \ \forall \mu$ (Q.E.D)
Algorithm 1 : Value iteration (VI)

According to fixed-point thorem, we have

\[ R_{\mu^*} = \lim_{N \to \infty} T \circ \ldots \circ T(R), \forall R \]

Let \( \{R_n\} \) with any function \( R_0 \) s.t.

\[ R_{n+1} = T(R_n) \]
\[ R_{n+1}(x) = \max_a \left\{ r(x, a) + \alpha \int R_n(y)Q(y|x, a)dy \right\} \]
\[ \mu_n(x) = \arg \max_a \left\{ r(x, a) + \alpha \int R_n(y)Q(y|x, a)dy \right\} \]

**Theorem**

\[ \lim_{N \to \infty} \mu_N(x) = \mu^*(x), \forall x \]
Algorithm 2: Linear Programming (LP)

- Finite set of states $\mathcal{X} = \{x^{(0)}, \ldots, x^{(V)}\}$
- Finite set of actions $\mathcal{A} = \{a^{(0)}, \ldots, a^{(V')}\}$

**Lemma**

Let $T$ be the Bellman’s operator on vector s.t. $\tilde{W} = T(W)$

$$\tilde{W}(m) = \max_a \left\{ r(x^{(m)}, a) + \alpha \sum_{n=0}^{V} W(n) Q(x^{(n)}|x^{(m)}, a) \right\}$$

Let $\geq$ be the elementwise “greater than”.

- If $U \geq V$, then $T(U) \geq T(V)$
- Let $W^*$ be fixed point of $T$ and $W$ s.t. $W \geq T(W)$, then $W \geq W^*$
- As $W^* \geq T(W^*)$, we get

$$W^* = \arg \min_W \sum_{m=0}^{V} W(m) \text{ s.t. } W \geq T(W)$$
Algorithm 2: Linear Programming (LP) (cont’d)

Function \( R_{\mu^*} \iff \text{Vector } \mathbf{R}^* = [R_{\mu^*}(x^{(0)}), \ldots, R_{\mu^*}(x^{(V)})] \)
\( \iff \mathbf{R}^* \text{ fixed point of } T \)

Linear programming algorithm

\[
\mathbf{R}^* = \arg\min_{\mathbf{R}} \sum_{m=0}^{V} R(n)
\]

s.t.

\[
R \geq T(R) \iff R(m) \geq \max_a \left\{ r(x^{(m)}, a) + \alpha \sum_{n=0}^{V} R(n)Q(x^{(n)}|x^{(m)}, a) \right\}
\]

i.e.

\[
R(m) \geq r(x^{(m)}, a^{(t)}) + \alpha \sum_{n=0}^{V} R(n)Q(x^{(n)}|x^{(m)}, a^{(t)}), \forall m, t
\]
Extension: CMDP

- Deterministic stationary policy not optimal anymore
- It exists a random stationary policy
- Computation of optimal policy still through linear programming
Part 3 : Applications to wireless networks
Multiband Exploration [Lun15]

- One secondary user may use $N_c$ non-contiguous channels (in 4G with carrier aggregation)
- Can explore only $N_e$ channels simultaneously

**Issue**: which ones selecting at any time?

**Partially Observable Markov Decision Process (POMDP)**

- **States** (at time $n$) - not all known - ... empty/occupied channels
- **Action** .................. $N_e$ channels to explore
- **Transition kernel** $Q(s_{n+1}|s_n, a_n)$  .................. Markovian process
- **Reward** $r(s_n, a_n)$  .......................... $\#$empty tested channel
- **Policy** $a_n = \pi(s_n)$  ..........................

Policy maximizing $\mathbb{E} \left[ \sum_n \alpha^n r(s_n, a_n) \right]$

**Extension**: Competition between secondary users $\Rightarrow$ stochastic game
Hybrid ARQ optimization

- **Type-I HARQ**: Let $S$ be a packet composed by $N$ coded symbols.

  ![Type-I HARQ Diagram]

  - $Y_1 = S_1(1) + N_1$
  - $Y_2 = S_1(2) + N_2$  $\Rightarrow$ detection on $Y = [Y_1, Y_2]$ (Coding gain)

- **Type-II HARQ**: Memory at RX and non-identical packets at TX.

  ![Type-II HARQ Diagram]
Case 1: Power optimization [Taj13]

We would like to adapt the power packet per packet
After the $n$-th transmission, if NACK is received, we have

- New action $A_{n+1}$ to do: here choosing $P_{n+1}$
- Available information: accumulated mutual information equal to

$$I_n = \sum_{\ell=1}^{n} \log_2(1 + G_\ell P_\ell)$$

- $K_n \in \{\text{ACK and new round, 1 attempt and NACK received,} \ldots, \text{L attempts and NACK received}\}$

$$P(K_{n+1}, I_{n+1}|K_n, I_n, \ldots) = P(K_{n+1}, I_{n+1}|K_n, I_n)$$

$\Rightarrow$ Markov Chain: $S_n = (K_n, I_n)$

A policy $\mu = \text{how selecting } P_{n+1} \text{ given } S_n$
Optimization problem and numerical results

\[
\max_{\mu} \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \text{reward}_\mu(S_n, A_n)
\]

\[
\text{s.t.}
\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \text{power}(S_n, A_n) \leq P_{\text{max}}
\]

- Optimal random policy exists
- Optimal pdf obtained through linear programming

---

**SNR (dB) [source: Tajan's PhD]**

**Throughput (in bcpu)**

- Constant power
- Optimal random policy based power
Case 2: Best Modulation and Coding scheme [Ksa15]

- Action: Best modulation and coding for the first packet transmission
- Action: Best modulation for the retransmission
  - if BPSK: a few redundancy bits sent but well protected
  - if QAM: a lot of redundancy bits sent but not well protected
- State: current channel impulse response, number of HARQ attempts, effective SNR

Numerical result

- LTE set-up
- Correlated channel (25km/h)
Scheduling with energy harvesting [Faw17]

**Action**: $u_n$ sent packets (of length $L$) using the following energy

$$C_n = \frac{\sigma^2}{g_n} (2^{u_nL} - 1)$$

**States**: $s_n$

- $k_n$: age of the packet within the buffer
- $b_n$: battery level $\rightarrow b_{n+1} = \min(b_n - C_n + e_{n+1}, B)$
- $g_n$: channel gain
**Examples MDP Applications**

**Exploration HARQ optimization Scheduling**

**Scheduling with energy harvesting (cont’d)**

**Reward**: packet loss (buffer overflow, delay non-fulfillment)

\[
\min_{N \to \infty} \lim_{N \to \infty} \mathbb{E} \left[ \frac{1}{N} \sum_{n=0}^{N} (\epsilon_o(s_n, u_n) + \epsilon_d(s_n, u_n)) \right]
\]

with

- \(\epsilon_o\) number of dropped packets due to buffer overflow
- \(\epsilon_d\) number of dropped packets due to delay non-fulfillment

**Numerical results**

[Graph showing dropped packet probability versus \(\lambda\) for different values of \(\lambda_e\)]
Not treated issues

- Curse of dimensionality
- Competition between users (game theory [Lun15])
- Unknown $Q$ : reinforcement learning, Q-learning
- Closely related to Dynamic Programming (Viterbi algorithm for channel decoding or ISI management)