Fast reinforcement learning for energy-efficient wireless communications

INF8225 – Project

Farnoush Farhadi   Juliette Tibayrenc

Polytechnique Montreal

April 2016

- Practical application of Markov Decision Processes (MDPs)
- Explore algorithm variants and compare results

Contents

Contents

Context

- several ways to optimize power consumption while transmitting delay-sensitive information
  - on the software side...
  - and on the hardware side
- but no strategy to ally both
- Unknown dynamic environments
  - Dynamic traffic and channel conditions
  - Lack of statistical knowledge of dynamics
  - Fast learning algorithms
- Heterogeneous multimedia data
  - Different deadlines, priorities, dependencies

Contents

Problem

Find a way to solve the optimization problem (balancing the constraints of low power consumption and low transmission time)

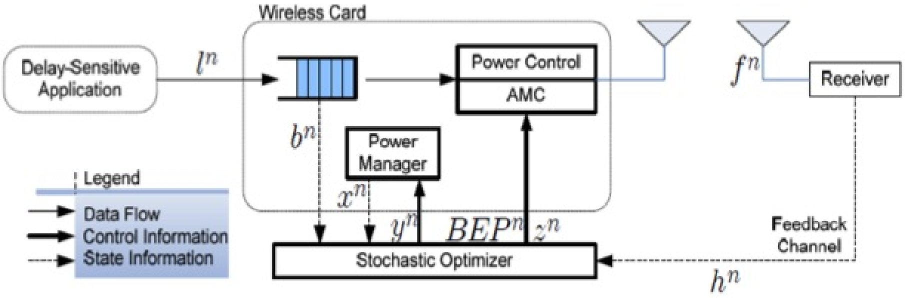

Figure: Power consumption minimization s.t buffer delay constraint (from the original article)

Rule: Average buffer delay is proportional to average buffer occupancy

Contents

Proposed solution

- Power management problem $\equiv$ MDP
- Separate known and unknown components (generalize* the PDS concept)
- Use reinforcement learning to solve the DPM problem

Contents

Related algorithms

- Value iteration and policy iteration
  - iterative algorithms
  - aim: find the optimal policy (directly or indirectly by optimizing the value function)
- Reinforcement learning & Q-learning
  - dynamics & reward function initially unknown
  - Q-learning: learn the best policy from history

A few results

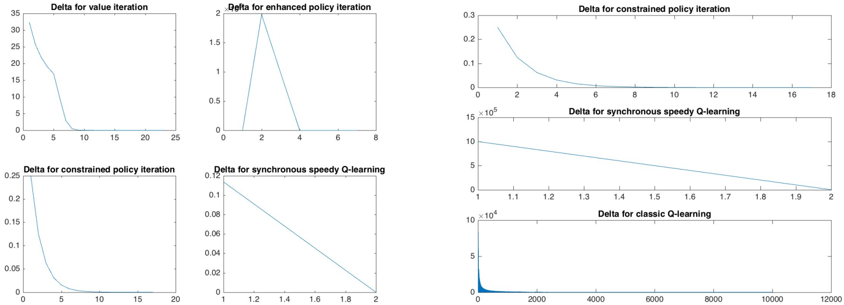

Figure: Comparing the required number of iterations to get to the optimal policy

Caution: not the measure that's the most indicative of performance

Contents

Reward matrix

- Reward obtained for getting to one state?
- $\Longrightarrow$ algorithm to compute the reward matrix

Algorithm 1 The Buffer State Transition Matrix Algorithm

Inputs:

$n \in \{0, 1, ...\}$

l^n : `Data arrival, i.i.d`

$$\text{states} = \begin{cases} b^n \in \{0, \cdots, \mathrm{B}\} : \texttt{Buffer State} \\ h^n : \texttt{Channel State} \\ x^n \in \{\mathrm{on}, \mathrm{off}\} : \texttt{Power Management State} \end{cases}$$

$$\text{actions} = \begin{cases} z^n, 0 \le z^n \le b^n : \texttt{Packet Throughput} \\ \mathrm{BEP}^n : \texttt{Bit Error Probability} \\ y^n \in \{s_{\mathrm{on}}, s_{\mathrm{off}}\} : \texttt{Power Management Action} \\ f^n, 0 \le f^n \le z^n : \texttt{Goodput} \end{cases}$$

Initialize:

$b^0 \leftarrow b_{init}$

$b^n \leftarrow \min(b^n - f^n(\mathrm{BEP}^n, z^n) + l^n, \mathrm{B})$

$\mathrm{P}^x = \mathrm{P}^x(y) = [\mathrm{P}^x(x^{'}|x, y)]_{x,x^{'}}$

$\mathrm{P}^h = \mathrm{P}^h(h^{'}|h)$

$$\mathrm{P}^b = \mathrm{P}^b([b, h, x], \mathrm{BEP}, y, x) = \begin{cases} \sum_{f=0}^{z} \mathrm{P}^l(b^{'} - [b - f])\mathrm{P}^f(f|\mathrm{BEP}, z), & \text{if } b^{'} \le \mathrm{B} \\ \sum_{f=0}^{z} \sum_{l=\mathrm{B}-[b-f]}^{\infty} \mathrm{P}^l(l)\mathrm{P}^f(f|\mathrm{BEP}, z), & \text{if } b^{'} = \mathrm{B} \end{cases}$$

$s \leftarrow (b, h, x)$

$a \leftarrow (\mathrm{BEP}, y, z)$

$\mathrm{P}(s^{'}|s, a) = \mathrm{P}^b \times \mathrm{P}^h \times \mathrm{P}^x$

Conclusion

The authors:

- Considered the problem of energy-efficient point-to-point transmission of delay sensitive over a fading channel.
- Proposed a unified reinforcement learning solution for finding the jointly optimal power-control, AMC, and DPM policies when the traffic arrival and channel statistics are unknown.
- Exploited the structure of the problem:
  - introducing a post-decision state
  - eliminating action-exploration
  - enabling virtual experience to improve performance

Conclusion

We:

- synthetised this work & reproduced results
- focused on the reward matrix subproblem
- implemented other approaches to solve the problem

Proposed algorithm outperforms existing solutions; our first algorithm outperforms it on some measures but don't present the same advantages.

Possibilities for a future work

Can be applied to any network or system resource management problem involving controlled buffers. $\Longrightarrow$ Apply it in a system with multiple users by integrating the single-user optimization with one of the multi-user resource allocation framework (ex: uplink or downlink transmission in cellular systems)

References

1. Mastronarde, N. and Van der Schaar, M., 2011. Fast reinforcement learning for energy-efficient wireless communication. Signal Processing, IEEE Transactions on, 59(12), pp.6262-6266.
2. Wiering, M. and Schmidhuber, J., 1998. Fast online Q (). Machine Learning, 33(1), pp.105-115.
3. Ghavamzadeh, M., Kappen, H.J., Azar, M.G. and Munos, R., 2011. Speedy Q-learning. In Advances in neural information processing systems (pp. 2411-2419).