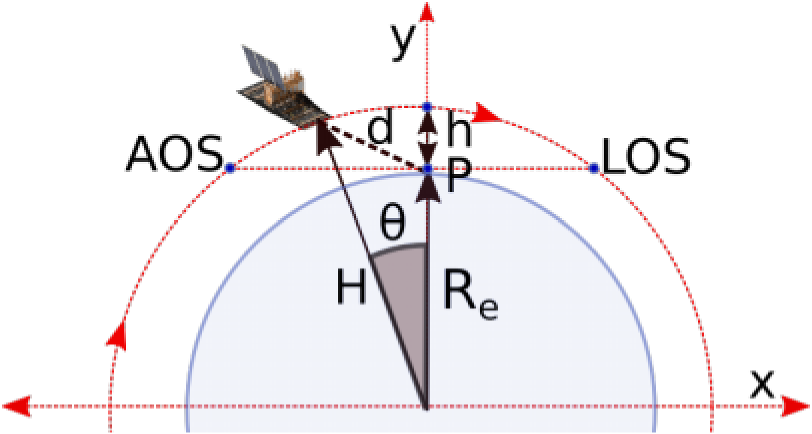
Adaptive Modulation Using Multi-Objective Reinforcement Learning for LEO Satellites

Graciela Corral Briones, Martín Ayarde, Adrián Ramirez and Felipe Pasquevich

UNC- CONAE - Argentina

21-23 June 2021

Geometrical representation of the LEO satellite orbit.



Equations

$$\theta(t) = \frac{2\pi t}{T} + \theta_i \tag{1}$$

$$T = 2\pi\sqrt{\frac{H^3}{\mu}}, \tag{2}$$

$$d = \sqrt{x^2 + (y - R_e)^2} \tag{3}$$

$$x(t) = H\sin(\theta(t)) \tag{4}$$

$$y(t) = H\cos(\theta(t)) \tag{5}$$

Comunication link

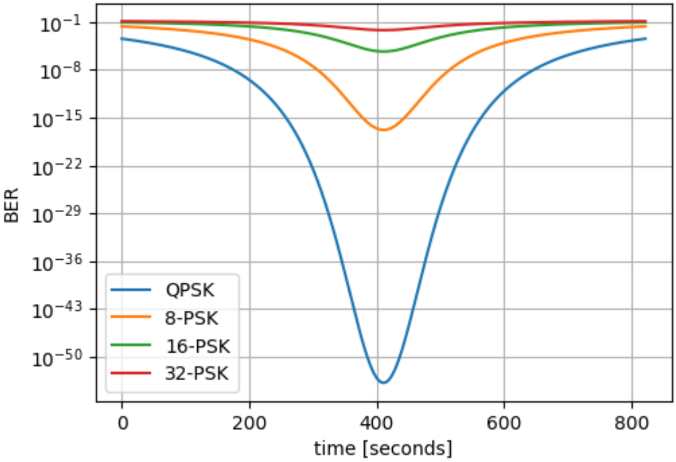
$$\text{FSPL} = 20\,log\left(\frac{4\,\pi\,d\,f}{c}\right) \tag{6}$$

$$\frac{C}{N_o} = \text{EIRP} - \text{FSPL} + \frac{G}{T} - 10\log\left(k\right) \tag{7}$$

$$\frac{E_b}{N_o} = 10^{\left(\frac{C}{N_o} - 10\log\left(R_b\right)\right)/10} \tag{8}$$

$$BER \approx \frac{2}{N}Q\left(\sqrt{2N\frac{E_b}{N_o}}sin\left(\frac{\pi}{2^N}\right)\right). \tag{9}$$

BER for a LEO GS downlink with a line of sight communication.

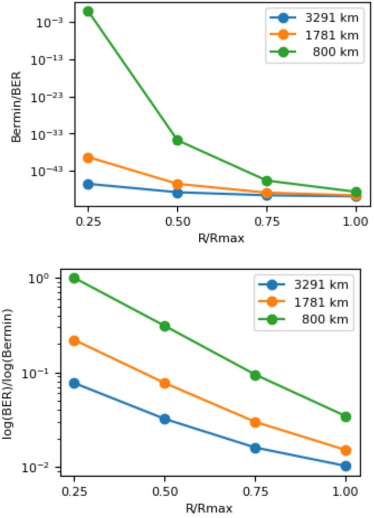


Simulation parameters

Table 1: Simulation parameters

Parameter	Abbr.	Value	Unit
LEO Altitude	h	639	km
Effective isotropic radiated power	EIRP	16	dBm
Antenna gain to noise temperature	G/T	24.83	dB/K
Carrier frequency	f	$2255\,10^6$	Hz
Symbol rate	R_s	$1\,10^6$	symb/sec
Simulation time step	Δt	8	sec

Normalization - Dynamic range



Weighted sum approach

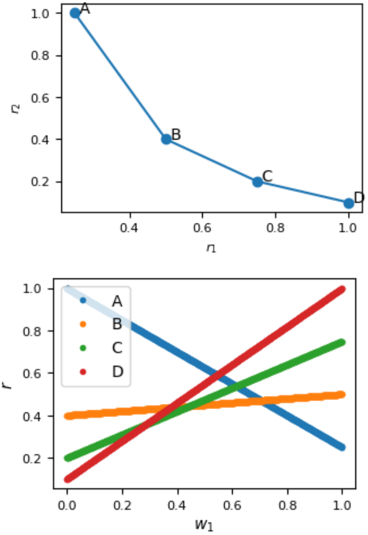
Scalarization:

$$r = r_1 w_1 + r_2 w_2 \tag{10}$$

conditioned to :

$$w_1 + w_2 = 1 \tag{11}$$

Pareto front

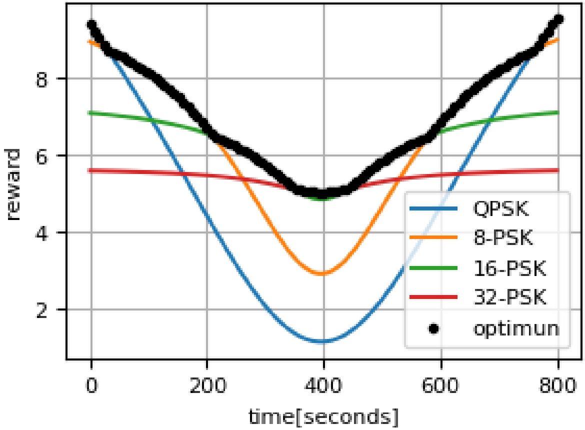


Inverse Scalarization

To avoid the weighted sum limitation we can define a new scalarization function that computes the inverse of the weighted sum.

$$r(r_1, r_2) = \frac{1}{r_1 w_1 + r_2 w_2} \quad \text{and} \quad w_1 + w_2 = 1 \tag{12}$$

Inverse scalarization function at time interval from AOS to LOS. The maximum r_{max} is displayed in the black curve.



Q-Learning

- Markov Decision Process

$$Q(s_t, a_t) \leftarrow \quad Q(s_t, a_t)(1 - \alpha)[r_{t+1} + \gamma \max_a Q(s_{t+1}, a)] \quad (13)$$

- Multi-armed Bandit Problem

$$Q(a_t) \leftarrow \quad Q(a_t)(1 - \alpha) + \alpha[r_{t+1} + \gamma \max_a Q(a)] \quad (14)$$

$s = \text{Satellite position}$
$a = \{QPSK, 8PSK, 16PSK, 32PSK\}$
$r = \frac{1}{r_1 w_1 + r_2 w_2} \quad \text{and} \quad w_1 + w_2 = 1$

Episodes, Returns and Policy

- Episodes
  $N = 823$ time steps, beginning at AOS and ending at LOS
- Returns

$$G = \sum_{k=0}^{N} r_k \tag{15}$$

$$G_r = \frac{G - G_{min}}{G_{max} - G_{min}} \tag{16}$$

- Policy
  $\epsilon - greedy$

Value function

$$\mathrm{CQ}(s,a)_j = \min(Q(s,a)_j, C_j) \tag{17}$$

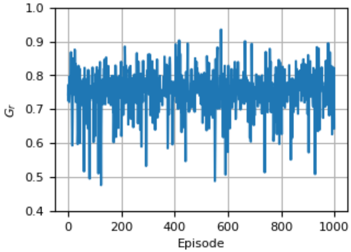
Function:

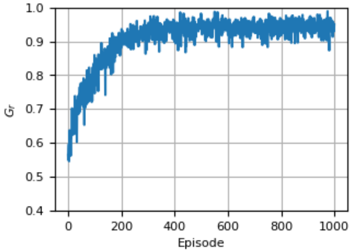
$\mathrm{Superior}(\mathrm{CQ}(s,a'), \mathrm{CQ}(s,a), 1)$
a' is true for all the actions a.

Python script

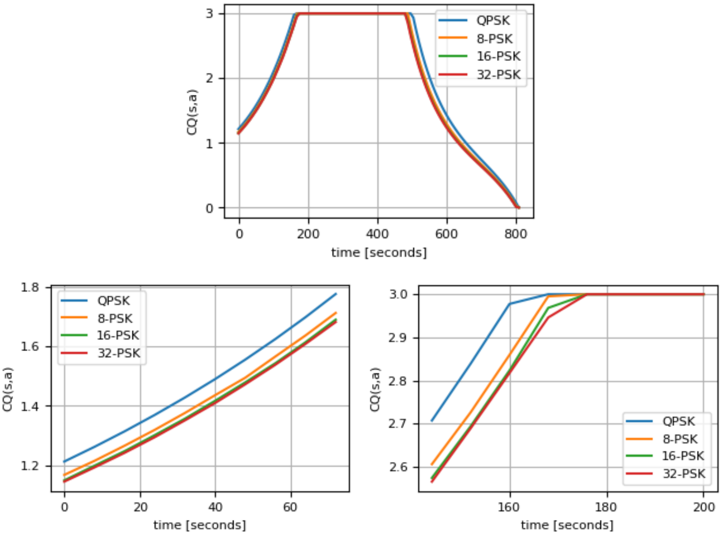
```
def Superior(CQ,i):
    if CQ[2*(i-1)] > CQ[2*i-1]:
        return True
    elif CQ[2*(i-1)] == CQ[2*i-1]:
        if i==numberOfObjectives:
            return True
        else:
            return Superior(CQ,i+1)
    else:
        return False
```

Returns Multi-armed Bandit and Markov Decisión Proccess





Threshold Lexicographic Q-Learning



Conclusions

- After analytically examining the weighted sum approach, we see it is not useful for our proposed case because the intermediate solutions of the objective pairs are not found, as a consequence of the non-convex Pareto front constraint.
- By performing scalarization transformations and assigning the appropriate reward to the learning agent by choosing weights, the inverse scalarization function allows us to obtain the solutions for the four digital modulation techniques.
- On the other hand, the TLQ algorithm starts by choosing the best modulation prioritizing the BER (i.e., QPSK) and once the threshold is reached, it randomly chooses the next modulation, until returning back to the previous same.