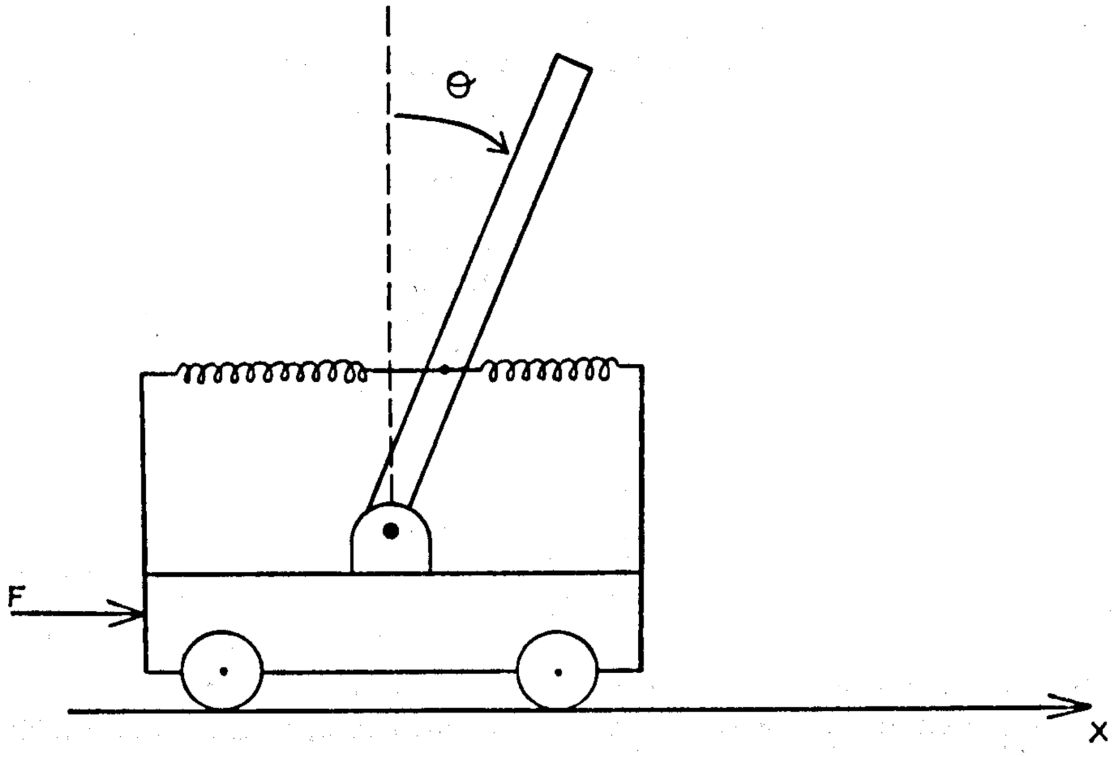
PRÁCTICA III

Aprendizaxe por reforzo

Pablo Chantada Saborido & José Romero Conde



1. Introdución e aspectos xerais

Sobre a programación, comentar:

2. Hiperparámetros

¿Qué hiperparámetros has utilizado? ¿Cómo has seleccionado estos hiperparámetros? ¿Por qué son más adecuados que otros valores?

3. Mellor política determinista

4. Mellor algoritmo de control

5. Perturbacións

6. Conclusións

For the control problem (finding an optimal policy), DP, TD, and Monte Carlo meth- ods all use some variation of generalized policy iteration (GPI). The differences in the methods are primarily differences in their approaches to the prediction problem.

Although Q- learning actually learns the values of the optimal policy, its on-line performance is worse than that of Sarsa, which learns the roundabout policy. Of course, if were gradually reduced, then both methods would asymptotically converge to the optimal policy.

Referencias