## Título del trabajo

“Estudio y desarrollo de algoritmos de aprendizaje por refuerzo a partir de la teoría de optimización dual”

“Reinforcement learning algorithms from duality theory”

## Nombre del tutor

Santiago Zazo Bello <santiago.zazo@upm.es>

## Resumen

Uno de los problemas clave en aprendizaje automático es la toma de decisiones para conseguir un objetivo. Este tipo de problemas se suele denominar "sequential learning" y uno de los marcos de trabajo más prometedores es el denominado aprendizaje por refuerzo o "reinforcement learning". Los algoritmos de reinforcement learning han sido motivados principalmente estableciendo conexiones con la teoría de control óptimo.

En la actualidad existen diversos algoritmos de aprendizaje por refuerzo que implementan programación dinámica estimando la función valor del problema de forma estocástica, es decir, a partir de muestras y sin conocimiento previo de los datos del problema. Aunque dichos algoritmos son completamente válidos, es interesante buscar nuevas interpretaciones que den lugar a nuevos algoritmos.

El objetivo de este TFM es el estudio de algoritmos de aprendizaje por refuerzo desde el punto de vista de la teoría de optimización. Para ello se llevará a cabo una revisión de los fundamentos de optimización convexa, teoría de la dualidad y teoría de control óptimo. Finalmente, se estudiará el estado del arte en algoritmos de aprendizaje por refuerzo.

Una vez completada la fase de revisión de estas teorías, se investigará la conexión entre el problema de control óptimo y su formulación como un problema de programación lineal, con la intención de desarrollar nuevos algoritmos de optimización estocástica que resuelvan el problema de aprendizaje por refuerzo. Asimismo, se realizarán diversas pruebas de convergencia y comparación de resultados con otros algoritmos ya existentes con el objetivo de evaluar la utilidad e idoneidad de nuestros algoritmos.

Metodología

La metodología a seguir a lo largo de este trabajo será:

- La formación de una base teórica sólida que constará de: (1) una revisión de los fundamentos de optimización convexa (conjuntos y funciones convexas, problemas de optimización convexa, teoría dual y problema dual de Lagrange, etc…), (2) el estudio de la teoría de control óptimo (formulación de programas lineales para resolver problemas de optimización; minimización del coste total) y (3) el estudio de algoritmos de aprendizaje por refuerzo (toma de contacto con el problema de aprendizaje por refuerzo, programación dinámica, métodos de Monte Carlo, métodos de predicción “Temporal-Difference” y métodos de control “SARSA” y “Q-learning”, algoritmos con aproximación de funciones como “LSPI” y de búsqueda en el espacio de políticas con métodos “policy gradien” y “actor-critic”).

- Formulación del problema de control óptimo como un problema saddle point del Lagrangiano: se realizará una propuesta de métodos para encontrar el saddle point a partir de las condiciones KKT con algoritmos primal-dual, así como otras formulaciones tipo dual-ascent o similares. Se plantearán algoritmos basados en modelo y también sus versiones estocásticas, con el objetivo de resolver el problema de control cuando no hay un modelo disponible.

- Evaluación de los algoritmos propuestos mediante simulación en problemas sencillos que sirvan como test (variaciones de random walk y grid-worlds) y comparación con algoritmos de referencia disponibles como “SARSA” y “Q-learning”.

- Extensión de los algoritmos propuestos al caso de aproximación paramétrica de la función valor. Comparación con algoritmo LSPI en problemas test estándar.

- Estudio de vínculos entre los algoritmos propuestos con otros algoritmos existentes de búsqueda en el espacio de políticas, en particular con policy gradient y métodos actor-crític.

- Análisis de convergencia mediante simulación y/o teórico de los algoritmos propuestos.

En lo que respecta a las herramientas de simulación a emplear, se hará uso de lenguajes de computación científica tales como Matlab o Python.

# Referencias

|  |  |
| --- | --- |
| [1] | R. S. Sutton y A. G. Barto, Reinforcement Learning: an introduction, The MIT Press, 2015. |
| [2] | S. Boyd y L. Vandenberghe, Convex Optimization, Cambridge University Press. |
| [3] | D. P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, 2012. |
| [4] | T. Wang, D. Lizotte, M. Bowling y D. Schuurmans, Dual Representations for Dynamic Programming, Journal of Machine Learning, 2008. |