# Tareas desarrolladas (incluyendo las fechas en que se han realizado)

A lo largo de este período (01.02.2017 – 24.03.2017) se ha llevado a cabo una revisión de los fundamentos de optimización convexa, teoría de la dualidad y teoría de control óptimo, que tendrá como objetivo final el desarrollo de nuevos algoritmos de aprendizaje por refuerzo. De este modo, los hitos que se fueron marcando fueron los siguientes:

A.1) Introducción al campo del aprendizaje por refuerzo: a través de [1] se llevó a cabo la primera toma de contacto con los conceptos relativos al reinforcement learning o aprendizaje por refuerzo (formulación del problema de aprendizaje por refuerzo, programación dinámica (policy y value iteration), métodos de Monte Carlo, métodos de predicción “Temporal-Difference” y métodos de control “SARSA” y “Q-learning”, etc…). Con el objetivo de asentar mejor las bases de estos métodos, se siguió además el video-curso online “Reinforcement Learning” proporcionado por el “Georgia Institute of Technology”. Al mismo tiempo que se avanzaba con los conocimientos teóricos, se realizaron una serie de ejercicios que aparecen propuestos en [1] para comprobar el correcto entendimiento de todos los conceptos. (**Fin hito A.1: 01.02.2017 – 15.02.2017**)

A.2) Introducción a los algoritmos existentes de aprendizaje por refuerzo: con [2] se tomó una perspectiva más completa del aprendizaje por refuerzo descubriendo la importancia que toman el operador de Bellman y las cadenas de Markov. De este modo, se trata de una interpretación más técnica de la que se encontró en [1], con la que además tuve un primer acercamiento al tipo de algoritmos con los que se trabajará. Para completar este enfoque, se empleó además [3], donde se sintetizan los conceptos de la programación dinámica y el aprendizaje por refuerzo y se muestran además implementaciones de los algoritmos más empleados a día de hoy: SARSA y Q-learning. Con el objetivo de comprobar el correcto entendimiento de toda la literatura leída tanto en el hito A.1 como en el A.2, se llevó a cabo la implementación de los dos algoritmos anteriormente citados, de manera que resolviesen 3 problemas tipo que aparecen en [1]: gridworld, cliff problem y random walk. (**Fin hito A.2: 15.02.2017 – 18.02.2017**)

Con la realización de ejercicios citada en el hito A.1 así como con la implementación de los algoritmos SARSA y Q-learning para algunos problemas tipo, se dio por finalizada la primera etapa de aprendizaje de los fundamentos básicos sobre aprendizaje por refuerzo.

B.1) Revisión de los fundamentos de optimización convexa: a través de [4], se formó una base teórica sólida en lo que a los conceptos de convexidad se refiere (conjuntos y funciones convexas, problemas de optimización convexa, teoría dual y problema dual de Lagrange, etc…). Para afianzar los conocimientos obtenidos, se siguió además el video-curso online “Convex Optimization” proporcionado por la universidad de Stanford impartido por el profesor Stephen Boyd, autor de [4]. Como se verá más adelante, en este primer hito cobrará especial relevancia la teoría dual. De nuevo, al mismo tiempo que se avanzaba con los conocimientos teóricos, se realizaron una serie de ejercicios que aparecen propuestos en [4] para comprobar el correcto entendimiento de todos los conceptos. (**Fin hito B.1: 18.02.2017 – 04.03.2017**)

Teniendo ya una base teórica apropiada en aprendizaje por refuerzo y optimización convexa, se pasó a una literatura que combina ambos conceptos orientándolos a la teoría del control óptimo:

C.1) Estudio de la conexión entre el problema de control óptimo y su formulación como un problema de programación lineal: en [5] se presenta de una manera más formal y técnica la formulación de programas lineales para resolver problemas de optimización con el objetivo de minimizar un coste total. Este enfoque se extiende al aprendizaje por refuerzo a través de value iteration (o enfoque de aproximaciones sucesivas) haciendo uso del operador de Bellman. (**Fin hito C.1: 05.03.2017 – 09.03.2017**)

Llegados a este punto se pudo notar que nuestro problema de aprendizaje por refuerzo se puede formular, bien como un programa dinámico o bien como un programa lineal. En principio, la visión que nos va a interesar es la del programa dinámico ya que dará pie a algoritmos de aproximación estocástica que permitirán mejorar con la experiencia generada conforme avanza el tiempo. No obstante, en la actualidad SARSA y Q-learning suponen algoritmos muy aceptados en la comunidad del aprendizaje por refuerzo que permiten una implementación estocástica.

El trabajo de investigación trata, por tanto, de la búsqueda de nuevos algoritmos que sean capaces de explotar la experiencia generada a lo largo del tiempo, haciendo uso de la formulación como un programa lineal que se expone en [5]. Y es en este punto donde toma especial relevancia la teoría dual mencionada en el hito B.1

D.1) Búsqueda de algoritmos con un enfoque primal-dual que permitan la reformulación del programa lineal expuesto en [5] para el problema de optimización de la función de coste (value function): a través de [6] se estudió la idea de formular el problema de control óptimo como un problema saddle point del Lagrangiano, y es en este punto donde comienzan las aportaciones de toda la formación expuesta en los hitos anteriores. Para abordar este problema de saddle point, se propuso el algoritmo Arrow-Hurwicz, que tras varias pruebas y consultas a diversas referencias se pudo comprobar que no era válido para nuestro caso. A continuación, se pasó al enfoque dual-ascent, con el cual sí se encontraron pruebas de convergencia, aún pendientes de un testeo completo y demostración matemática completa para nuestro problema en cuestión.

Con este nuevo enfoque basado en la teoría dual, se demuestra en [6] que de la variable dual se puede extraer la política de comportamiento del problema, con lo cual, resolver el problema de saddle point encontrando la variable dual óptima supone encontrar la política óptima. Es decir, el nuevo método dual planteado permite buscar la solución óptima en el espacio de políticas en lugar de en el espacio de las funciones valor tal y como hacen los conocidos algoritmos SARSA y Q-learning.

En vista de la aparente idoneidad de este nuevo enfoque desarrollado, 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. Actualmente se están probando estas nuevas ideas con lo que parecen buenos resultados de convergencia en comparación con sus competidores SARSA y Q-learning. No obstante, aún no se pueden sacar conclusiones relevantes. (**Fin hito D.1: 09.03.2017 – 24.03.2017**)

En este punto finalizan las tareas desarrolladas hasta la fecha actual.

# Referencias

|  |  |
| --- | --- |
| [1] | R. S. Sutton y A. G. Barto, Reinforcement Learning: an introduction, The MIT Press, 2015. |
| [2] | C. Szepesvari, Algorithms for Reinforcement Learning, Morgan & Claypool Publishers, 2009. |
| [3] | L. Busoniu, R. Babuska, B. De Schutter y D. Ernst, Reinforcement learning and dynamic programming using function approximators. |
| [4] | S. Boyd y L. Vandenberghe, Convex Optimization, Cambridge University Press. |
| [5] | D. P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, 2012. |
| [6] | T. Wang, D. Lizotte, M. Bowling y D. Schuurmans, Dual Representations for Dynamic Programming, Journal of Machine Learning, 2008. |

# Relación de problemas planteados y el procedimiento seguido para su resolución

El principal problema planteado ha sido el desarrollo de un nuevo algoritmo orientado al aprendizaje por refuerzo que sea competitivo frente a los ya existentes y más conocidos, tales como SARSA, Q-learning, LSPI, DQN o NFQ entre los más destacables. Para ello, se pasó por el proceso natural de: (1) resolver el problema de control óptimo en problemas de pequeña escala cuando se conoce el modelo y cuando no se conoce el modelo, y (2) resolver el problema de control óptimo en problemas en los que el espacio de estados es muy grande o continuo

Para resolver este problema, se llevó a cabo un estudio del estado del arte en lo que a reinforcement learning se refiere (hitos citados anteriormente), de manera que se concluyó que un método alternativo podría ser buscar en el espacio de políticas en lugar de en el de las funciones valor. Tras consultar varias referencias, se adoptó una interpretación que permite resolver el problema primal de optimización de una función valor, a través de su formulación como un problema max-min de saddle point del Lagrangiano, en el cual la variable dual es la política a seguir (y la primal sigue siendo la función valor).

De este modo, se ha conseguido reformular el problema inicial expresado en forma de problema de optimización de una función valor, como un problema de optimización de la política a seguir (búsqueda en el espacio de políticas).

Para abordar la resolución del problema de saddle-point, se recurrió primero al método propuesto por Arrow y Hurwicz. Al llevarlo a la práctica nos encontramos con que dicho algoritmo nunca convergía a la solución del saddle point, sino que era altamente oscilante. Consultando en la publicación en la que Arrow y Hurwicz presentan este método, se encontró que cuando la función objetivo es lineal, el algoritmo es altamente oscilante y no converge a menos que el punto de partida elegido se encuentre en una bola cercana al saddle point. Debido a este inconveniente, este primer método fue dejado de lado. Como solución alternativa, se pasó a probar el método dual-ascent, el cual se pudo comprobar que sí convergía al saddle point.

En lo que refiere a los problemas que surgieron a lo largo de los hitos desarrollados en el punto anterior estos fueron de diverso carácter:

Problemas en la resolución de ejercicios de base: tras toda la adquisición de la base teórica, se llevaron a cabo una serie de ejercicios para comprobar el correcto entendimiento de los conceptos de mayor relevancia. En ocasiones alguna idea no me quedó completamente clara y en consecuencia no era capaz de resolver el problema en cuestión. De este modo, consultando a los compañeros del grupo fui capaz de aclarar todas las ideas y resolver los problemas de carácter teórico.

Problemas de implementación: en la implementación de algoritmos, no siempre se obtenía el resultado esperado. Para solucionar estos casos, se realizaba un proceso de debugging en el cual podía detectar errores conceptuales (aspectos teóricos que pensé que me habían quedado claros, pero no era así) así como de implementación. Me fue de gran utilidad la consulta de distintos foros que trataban estas temáticas en internet, así como la lectura de diversos artículos y más documentación que trataban aspectos de implementación.

# Valoración personal de la experiencia en términos de aprendizaje y competencias adquiridas

En el tiempo que ha transcurrido desde el comienzo de la beca mis conocimientos sobre reinforcement learning han aumentado enormemente y de manera muy satisfactoria, pues al inicio mi desconocimiento sobre este campo era prácticamente total. Considero que, aunque aún ha pasado un corto periodo de tiempo, he conseguido llevarme una visión global de todas las problemáticas existentes en el aprendizaje por refuerzo (aún muy ampliable esta visión) lo cual me va a dar pie a poder abordar diferentes tipos de problemas en un futuro, así como alcanzar resultados satisfactorios al término de la beca. Y este progreso no ha sido sólo debido a la lectura y puesta en práctica de toda la documentación expuesta en los hitos, sino que se ha visto favorecida por trabajar en un entorno de lo más “amigable” posible en el cual podía consultar al resto de miembros del grupo, y especialmente a Sergio Valcarcel, cualquier duda que me surgiese.

Al margen de la línea principal de trabajo, cada jueves se han realizado reuniones grupales en las que todos de miembros del grupo que hemos comenzado recientemente la labor de investigación exponíamos nuestro trabajo y nuestro progreso. De este modo, he adquirido nociones básicas sobre otras materias como son las redes neuronales, de gran utilidad en combinación con mi línea de investigación. Por tanto, en el tiempo que ha transcurrido de mi beca he podido nutrirme, no sólo del conocimiento que me proporciona mi trabajo, sino del generado por los demás.

Como competencia transversal, estoy aprendiendo a desenvolverme en la presentación de mis resultados en dichas reuniones, con lo que espero que mis habilidades de oratoria mejoren al final de la beca, y sobre todo mi vocabulario y expresividad en esta materia que es tan nueva para mí.

Por último y no menos importante, he aprendido (y estoy aprendiendo) lo que conlleva trabajar en el ámbito investigador. Acostumbrado durante los años de grado y máster a poner en práctica cosas que ya sabía que funcionaban de antemano, en esta beca he podido ver de cerca lo que supone abordar un nuevo problema desconocido y la tarea que ello implica: definir alternativas de testeo para asegurarse de que se avanza de manera correcta, apoyarse en demostraciones matemáticas que avalen el procedimiento a implementar. En definitiva, seguir una metodología de trabajo. En lo que respecta a lo personal, también supone enfrentarse a situaciones en las que todo el tiempo y trabajo invertido en algo que creía que iba a funcionar no sale como esperaba, de manera que aumenta la tolerancia a la frustración que suponen estas situaciones.

En resumen, creo que está siendo una experiencia muy enriquecedora en todos los aspectos (laborales, académicos y personales), ya que ha aumentado mi conocimiento sobre data science en general y reinforcement learning en particular, me está permitiendo mejorar mis habilidades referidas a la programación, tanto en Matlab como en Python, orientadas a reinforcement learning y me ha aportado un enfoque del trabajo investigador que desconocía hasta el momento así como la importancia de mantenerse actualizado al estado del arte.