

# Optimisme en apprentissage par renforcement et divergence de Kullback-Leibler

Sarah Filippi, Olivier Cappé, Aurelien Garivier

► **To cite this version:**

Sarah Filippi, Olivier Cappé, Aurelien Garivier. Optimisme en apprentissage par renforcement et divergence de Kullback-Leibler. Journées MAS et Journée en l'honneur de Jacques Neveu, Aug 2010, Talence, France. <inria-00510327>

**HAL Id: inria-00510327**

**<https://hal.inria.fr/inria-00510327>**

Submitted on 18 Aug 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journées MAS 2010, Bordeaux

Session : Apprentissage par renforcement

## **Optimisme en apprentissage par renforcement et divergence de Kullback-Leibler**

par **Sarah Filippi**, Olivier Cappé et Aurélien Garivier

We consider model-based reinforcement learning in finite Markov Decision Processes (MDPs), focussing on so-called optimistic strategies. Optimism is usually implemented by carrying out extended value iterations, under a constraint of consistency with the estimated model transition probabilities. In this paper, we strongly argue in favor of using the Kullback-Leibler (KL) divergence for this purpose. By studying the linear maximization problem under KL constraints, we provide an efficient algorithm for solving KL-optimistic extended value iteration. When implemented within the structure of UCRL2, the near-optimal method introduced by [Auer&al, 2009], this algorithm also achieves bounded regrets in the undiscounted case. We however provide some geometric arguments as well as a concrete illustration on a simulated example to explain the observed improved practical behavior, particularly when the MDP has reduced connectivity. To analyze this new algorithm, termed KL-UCRL, we also rely on recent deviation bounds for the KL divergence which compare favorably with the  $L_1$  deviation bounds used in previous works.

*Adresses :*

Sarah FILIPPI

LTCI, CNRS UMR 5141, Telecom ParisTech

Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France

E-mail : [sarah.filippi@telecom-paristech.fr](mailto:sarah.filippi@telecom-paristech.fr)

<<http://www.telecom-paristech.fr/>>

Olivier CAPPÉ

LTCI, CNRS UMR 5141, Telecom ParisTech

Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France

E-mail : [olivier.cappe@telecom-paristech.fr](mailto:olivier.cappe@telecom-paristech.fr)

<<http://www.telecom-paristech.fr/~cappe>>

Aurélien GARIVIER

LTCI, CNRS UMR 5141, Telecom ParisTech

Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France

E-mail : [aurelien.garivier@telecom-paristech.fr](mailto:aurelien.garivier@telecom-paristech.fr)

<<http://www.telecom-paristech.fr/~garivier>>

Session : Apprentissage par renforcement