Bayesian Quadratic Reinforcement Learning

Philipp Hennig
Cavendish Laboratory
University of Cambridge
CB3 0HE Cambridge, UK
ph347@cam.ac.uk

David H. Stern and Thore Graepel
Microsoft Research Ltd.
Roger Needham Building
CB3 0FB Cambridge, UK
[DStern|ThoreG]@microsoft.com

Abstract

The idea of using "optimistic" evaluations to guide exploration has been around for a long time, but has received renewed interest recently thanks to the development of bound-based guarantees [Kolter and Ng, 2009]. Asmuth et al. [2009] have recently presented a way to construct optimistic evaluations by merging sampled MDPs. Their BOSS algorithm is polynomial in the size of the state-action-space of the merged MDP, and thus a multiple of the size of the actual state space.

The success of BOSS raises the question whether it might be possible to avoid the sampling step and instead use an approximate parametrization of the belief over values. To test this idea, we construct an approximate Gaussian (i.e. log quadratic) joint posterior over the values of states under a given policy. Our derivation ignores the effect of future expected experience (in contrast to a fully Bayesian treatment as in the BEETLE algorithm [Poupart et al., 2006]).

On a discrete state-space environment described by a vector of expected rewards $\bar{\boldsymbol{r}}$ (i.e. the expected reward from the state-action pair $(s,a)_i$ is $\bar{r}_i$), transition dynamics described by a matrix $\boldsymbol{T}$ (i.e. $p(s_j|(s,a)_i) = T_{ij}$), and a policy $\boldsymbol{\Pi}$ (i.e. $p((s,a)_j|s_i) = \Pi_{ij}$), the values $\boldsymbol{q}$ of state-action pairs are determined by Bellman's equation $\boldsymbol{\Pi}(\boldsymbol{I} - \gamma \boldsymbol{T}\boldsymbol{\Pi})\boldsymbol{q} = \boldsymbol{\Pi}\boldsymbol{T}\bar{\boldsymbol{r}}$. Using, for example, conjugate prior inference with a Dirichlet prior on $\boldsymbol{T}$ and a Gaussian prior on $\bar{\boldsymbol{r}}$, leads to a belief over $\boldsymbol{q}$ which, while obviously not analytically Gaussian, can often be characterized reasonably well by a mean vector and covariance matrix.

We draw inspiration from the Expectation Propagation algorithm [Minka, 2001], using moment matching to construct an approximate Gaussian marginal. Unfortunately, the structure of Bellman's equation makes it difficult to perform moment matching directly on the marginal of $\boldsymbol{q}$. But it is straight-forward to do so on $\bar{\boldsymbol{r}}$, from which an approximate Gaussian message to $\boldsymbol{q}$ can be constructed. The resulting approximate belief over $\boldsymbol{q}$ provides optimistic point estimates as a function of the policy $\boldsymbol{\Pi}$, e.g. as $\|\boldsymbol{\mu}_{\boldsymbol{q}} + \beta \boldsymbol{e}_{\boldsymbol{q}}\|$, where $\boldsymbol{e}_{\boldsymbol{q}}$ is the eigenvector associated with the largest eigenvalue of the joint belief over $\boldsymbol{q}$.

We compared this approach to BEETLE and BOSS, on the widely used *Chain* environment. Our simplistic, parametric, approximate approach significantly outperforms the two more expensive contemporary algorithms on the most challenging, "full" setup, and shows an instructive disadvantage to them in the "semi-tied" setup. (The "tied" setup is too simple, all algorithms perform well on it). This suggests that there is scope for the application of light-weight probabilistic algorithms based on approximate message-passing, as solvers for general Reinforcement Learning Problems. Their modular structure, allows for the explicit modeling of environment structure and thus for increased performance on real-world problems.

References

John Asmuth, Lihong Li, Michael L. Littman, Ali Nouri, and David Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2009.

J.Z. Kolter and A.Y. Ng. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, 2009.

Thomas P. Minka. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2001.

Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd Annual International Conference on Machine Learning*. ACM New York, 2006.