the $p$-norm link function offers a flexible interpolation between additive and multiplicative gradient updates. It has been shown that when the features are dense and the optimal coefficients $\theta^*$ are sparse, EG converges faster than the regular additive gradient methods [KW95]. However, according to our experience, a significant drawback of EG is the overflow of the coefficients due to the exponential operator. To prevent overflow, the most commonly used technique is rescaling: the weights are re-normalized to sum to a constant. However, it seems that this approach does not always work. It has been pointed out [PS95] that in the EG-Sarsa algorithm, rescaling can fail, and replacing eligible traces instead of regular additive eligible traces is used to prevent overflow. EG-Sarsa usually poses restrictions on the basis as well. Thanks to the flexible interpolation capability between multiplicative and additive gradient updates, the $p$-norm link function is more robust and applicable to various basis functions, such as polynomial, radial basis function (RBF), Fourier basis [KOT08], proto-value functions (PVFs), etc.

### 7 Summary and Future Work

We proposed a novel framework for reinforcement learning using mirror-descent online convex optimization. Mirror Descent Q-learning demonstrates the following advantage over regular Q-learning: faster convergence rate and reduced variance due to larger stepsizes with theoretical convergence guarantees [NJLS09]. Compared with existing sparse reinforcement learning algorithms such as LARS-TD, Algorithm 2 has lower sample complexity and lower computation cost, advantages accrued from the first-order mirror descent framework combined with proximal mapping [SST11a]. There are many promising future research topics along this direction. We are currently exploring a mirror-descent fast-gradient RL method, which is both convergent off-policy and quicker than fast gradient TD methods such as GTD and TDC [SMP+09]. To scale to large MDPs, we are investigating hierarchical mirror-descent RL methods, in particular extending SMDP Q-learning. We are also undertaking a more detailed theoretical analysis of the mirror-descent RL framework, building on existing analysis of mirror-descent methods [DHS11, SST11a]. Two types of theoretical investigations are being explored: regret bounds of mirror-descent TD methods, extending previous results [SW94] and convergence analysis combining robust stochastic approximation [NJLS09] and RL theory [BT96, Bor08].

### Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-10-1-0383, and the National Science Foundation under Grant Nos. NSF CCF-1025120, IIS-0534999, and IIS-0803288. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the AFOSR or the NSF.

### References


