Combining No-regret and Q-learning

Ian A. Kash*
University of Illinois, Chicago, IL
Katja Hofmann
Microsoft Research, Cambridge, UK

Abstract

Most reinforcement learning algorithms do not provide guarantees in settings with multiple agents or partial observability. A notable exception is Counterfactual Regret Minimization (CFR), which provides both strong convergence guarantees and empirical results in settings like poker. We seek to understand how these guarantees could be achieved more broadly. To take a first step in this direction, we introduce a simple algorithm, local no-regret learning (LONR), which captures the spirit of CFR, but can be applied in settings without a terminal state. We prove its convergence for the basic case of MDPs and discuss research directions to extend our results to address richer settings with multiple agents, partial observability, and sampling.

Introduction

Reinforcement learning (RL) has seen significant successes in domains such as Atari games (Mnih et al. 2015) and robotics (Montgomery et al. 2017). A key feature shared by these domains is that they are single-agent and there is (close to) full observability of the environment. This has allowed these strong empirical results to be driven by (approximations of) algorithms with strong theoretical convergence guarantees. In contrast, in domains such as Go (Silver et al. 2016) and Doom (Jin, Levine, and Keutzer 2017) which lack at least one of these properties there has been empirical success but more limited theoretical justification for the algorithms.

An exception is poker, which has both multiple strategic agents and partial observability due to an inability to see opponents' cards (also known as *incomplete information*). Despite this, multiple algorithms have found success in playing at human expert level (Brown and Sandholm 2017; Moravčík et al. 2017) and non-trivial versions have been fully solved (Bowling et al. 2015). The algorithm underpinning these impressive results, Counterfactual Regret Minimization (CFR) (Zinkevich et al. 2008), is an algorithm developed for solving games of incomplete information. It works by using regret matching (a particular no-regret learning algorithm) to select actions. In particular, one copy of such an algorithm is used at each *information set*, which corresponds to the full history of play observed by a single agent. The resulting algorithm satisfies a global no-regret guarantee, so at least in two-player games is guaranteed to converge to an optimal strategy through sufficient self-play. Thus, this approach addresses both multiple agents and partial observability while having both theoretical guarantees and strong empirical results.

However, CFR does have limitations. It makes several strong assumptions which are natural for games of incomplete information such as poker, but limit applicability to further settings. For example, it assumes that the agent has perfect recall, which in a more general context means that the state representation captures the full history of states visited (and so imposes a tree structure). It also assumes that a terminal state is eventually reached and performs updates only after this occurs, which is not a requirement for traditional algorithms like Q-learning. Finally, it makes other specific assumptions, such as the use of a particular no-regret algorithm. Nevertheless, approaches inspired by CFR have shown empirical promise in domains that do not necessarily satisfy these requirements (Jin, Levine, and Keutzer 2017).

In this paper, we take a step toward putting this type of approach to general RL problems on a firmer theoretical foundation. We develop a new algorithm, which we call local no-regret learning (LONR), which in the same spirit as CFR uses a copy of an arbitrary no-regret algorithm in each state (for technical reasons we require a slightly stronger property we term no-absolute-regret). Our main result is that LONR has the same convergence guarantee as Q-learning for a Markov Decision Process (MDP). While our result does not immediately extend to multiple agents, partial observability, or sampling, we believe our result provides a starting point for progress on them and conclude with a discussion.

The closest technical approach to ours that we are aware of is the approach used by (Bellemare et al. 2016) to introduce new variants of the Q-learning operator. However, our algorithm is not an operator as the policy used to select actions changes from round to round in a history-dependent way, so we instead directly analyze the sequences of Q-values our algorithm generates. Additionally, unlike prior RL results but like prior no-regret learning results our proofs of convergence are for the limit of the average of the Q-

*This work was done while Ian Kash was a Researcher at Microsoft Research.

values rather than the Q-values themselves.

Related work

Beyond the work already discussed, the most closely related literature to our work is the literature on multi-agent learning. A common approach is to use no-regret learning as an outer loop to optimize over the space of policies, with the assumption that the inner loop of evaluating a policy is given to the algorithm. There is a large literature on this approach in normal form games (Greenwald and Jafari 2003), where policy evaluation is trivial, and a smaller one on "online MDPs" (Even-Dar, Kakade, and Mansour 2009; Mannor and Shimkin 2003; Yu, Mannor, and Shimkin 2009; Ma, Zhang, and Sugiyama 2015), where it is less so. Of particular note in this literature, (Even-Dar, Kakade, and Mansour 2005) also use the idea of having a copy of a no-regret algorithm for each state. An alternate approach to solving multi-agent MDPs is to use Q-learning as an outer loop with some other algorithm as an inner loop to determine the collective action chosen in the next state (Hu and Wellman 2003; Greenwald, Hall, and Serrano 2003). (Gondek, Greenwald, and Hall 2004) proposed the use of no-regret algorithms for this purpose. In contrast to these literatures, we combine RL in each step of the learning process rather than having one as an inner loop and the other as an outer loop.

Prior work has drawn other connections between no-regret learning and RL to reduce the sample complexity of Monte-Carlo Tree Search (Kaufmann and Koolen 2017), reduce imitation learning to no-regret learning (Ross, Gordon, and Bagnell 2011), and reduce RL to contextual bandits (Daumé III, Langford, and Sharaf 2018). There is also work on a RL algorithm which achieves good regret bounds (Jaksch, Ortner, and Auer 2010) and lower bounds on regret for RL algorithms (Osband and Van Roy 2016).

Preliminaries

Consider a Markov Decision Process $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the (finite) action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}$ is the transition probability kernel, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the (expected) reward function (which we assume to be bounded), and γ is the discount rate. In operator form, Q-learning is an operator $\mathcal{T}$ whose domain is bounded real-valued functions over $\mathcal{S} \times \mathcal{A}$ and is defined as

$$\mathcal{T}Q(s,a) = r(s,a) + \gamma \mathbb{E}_P[\max_{a' \in \mathcal{A}} Q(s',a')] \tag{1}$$

This operator is a contraction map in $||\cdot||_\infty$, and so converges to a unique fixed point Q^*, where $Q^*(s,a)$ gives the expected value of the MDP starting from state s, taking action a, and thereafter following the optimal policy $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$ (Bertsekas and Tsitsiklis 1996).

Our algorithm makes use of a no-regret learning algorithm. Consider the following (adversarial full-information) setting. There are n actions $a_1, \ldots a_n$. At each timestep k an online algorithm chooses a probability distribution π_k over the n actions. Then an adversary chooses a reward $x_{k,i}$ for each action i from some closed interval, e.g. $[0,1]$, which the algorithm then observes. The (external) regret of the algorithm at time k is

$$\frac{1}{k+1} \max_i \sum_{t=0}^{k} x_{t,i} - \pi_t \cdot x_t \tag{2}$$

An algorithm is *no-regret* if there a sequence of constants ρ_k such that regardless of the adversary the regret at time k is at most ρ_k and $\lim_{k\to\infty} \rho_k = 0$. For example, a common bound is that ρ_k is $O(1/\sqrt{k})$.

For our results, we make use of a stronger property, that the *absolute value* of the regret is bounded by ρ_k. We call such an algorithm a *no-absolute-regret* algorithm. Algorithms exist that satisfy the even stronger property that the regret is at most ρ_k and at least 0. Such *non-negative-regret* algorithms include all linear cost Regularized Follow the Leader algorithms, which includes Randomized Weighted Majority and linear cost Online Gradient Descent (Gofer and Mansour 2016).

Local no-regret learning

The idea of this work is to try and fuse the essence of Q-learning and CFR. A standard analysis of Q-learning proceeds by analyzing the sequence of matrices $Q, \mathcal{T}Q, \mathcal{T}^2Q, \mathcal{T}^3Q, \ldots$. The essence of CFR is to choose the policy for each state locally using a no-regret algorithm. While doing so does not yield an operator, as the policy changes each round in a history-dependent way, this process still yields a sequence of Q matrices as follows.

Fix a matrix Q_0. Initialize $|\mathcal{S}|$ copies of a no-absolute-regret algorithm with $n = |\mathcal{A}|$ and find the initial policy $\pi_0(s)$ for each state s. Then iteratively reveal the rewards to copy s of the algorithm as $x^s_{k,i} = Q_k(s, a_i)$, and update the policy π_{k+1} according to the no-absolute-regret algorithm and $Q_{k+1}(s,a) = r(s,a) + \gamma \mathbb{E}_{P,\pi_k}[Q_k(s',a')]$.

Call this process local no-regret learning (LONR). It can be viewed as a version of Expected SARSA (Van Seijen et al. 2009) where instead of using an ϵ-greedy policy with decaying ϵ, a no-absolute-regret policy is used instead. In the rest of this section we work up to our main result, that LONR converges to Q^*. Like many prior results using no-regret learning (e.g. (Zinkevich et al. 2008)), the convergence is of the average of the Q_k matrices.

We work up to this result through a series of lemmas. To begin, we derive a bound on the on average of Q values using the no-absolute-regret property. We need to use two slightly different averages to be able to relate them using the $\mathcal{T}$ operator.

Lemma 1. *Let* $\overline{Q}_k = 1/k \sum_{t=1}^{k} Q_t$ *and* $\underline{Q}_k = 1/k \sum_{t=0}^{k-1} Q_t$. *Then*

$$-\gamma\rho_{k-1} + \mathcal{T}\underline{Q}_k(s,a) \le \overline{Q}_k(s,a) \le \gamma\rho_{k-1} + \mathcal{T}\underline{Q}_k(s,a). \tag{3}$$

Proof. By the definitions of LONR and no-regret algo-

rithms,

$$\begin{aligned}
\overline{Q}_k(s,a) &= \frac{1}{k}\sum_{t=1}^{k} Q_t(s,a) \\
&= \frac{1}{k}\sum_{t=0}^{k-1} r(s,a) + \gamma \mathbb{E}_{P,\pi_t}[Q_t(s',a')] \\
&= r(s,a) + \gamma \mathbb{E}_P[\frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}_{\pi_t}[Q_t(s',a')]] \\
&\geq r(s,a) + \gamma \mathbb{E}_P[\max_i \frac{1}{k}\sum_{t=0}^{k-1} Q_t(s',a_i) - \rho_{k-1}] \\
&= -\gamma\rho_{k-1} + r(s,a) + \gamma \mathbb{E}_P[\max_i \frac{1}{k}\sum_{t=0}^{k-1} Q_t(s',a_i)] \\
&= -\gamma\rho_{k-1} + r(s,a) + \gamma \mathbb{E}_P[\max_i \underline{Q}_k(s',a_i)] \\
&= -\gamma\rho_{k-1} + \mathcal{T}\underline{Q}_k(s,a)
\end{aligned}$$

The key step is the inequality in the fourth line, where we use the fact that the policy for state s' is being determined by a no-regret algorithm, so we can use Equation (2) to bound the expected value of that policy terms of the value of the hindsight-optimal action and the regret bound of the algorithm. Similarly, by the stronger no-absolute-regret property, we can reverse the inequality to get $\overline{Q}_k(s,a) \leq \gamma\rho_{k-1} + \mathcal{T}\underline{Q}_k(s,a)$ This proves Equation (3). □

Next, we show that the range that the Q values take on is bounded. This lemma is similar in spirit to Lemma 2 of (Bellemare et al. 2016).

Lemma 2. *Let* $||r||_\infty = \max_{s,a}|r(s,a)|$. *Then* $||Q_k - Q_0||_\infty \leq 1/(1-\gamma)||r||_\infty + 2||Q_0||_\infty$

Proof. By definition, $Q_k(s,a) = r(s,a) + \gamma\mathbb{E}_{P,\pi_k}[Q_{k-1}(s',a')]$. Thus by the subadditive property of norms, $||Q_k||_\infty \leq ||r||_\infty + \gamma||Q_{k-1}||_\infty$. By induction, $||Q_k||_\infty \leq (\sum_{t=0}^{k-1}\gamma^k)||r||_\infty + \gamma^k||Q_0||_\infty$. Thus $||Q_k - Q_0||_\infty \leq ||Q_k||_\infty + ||Q_0||_\infty \leq 1/(1-\gamma)||r||_\infty + 2||Q_0||_\infty$. □

Combining these two lemmas, we can show that $\underline{Q}_k$ is an approximate fixed-point of $\mathcal{T}$, and that the approximation is converging to 0 as $k \to \infty$.

Lemma 3. $||\underline{Q}_k - \mathcal{T}\underline{Q}_k||_\infty \leq \frac{1}{k}(1/(1-\gamma)||r||_\infty + 2||Q_0||_\infty) + \gamma\rho_{k-1}$

Proof. Applying the bounds from Lemmas 1 and 2, we get that

$$\begin{aligned}
||\underline{Q}_k - \mathcal{T}\underline{Q}_k||_\infty &\leq ||\underline{Q}_k - \overline{Q}_k||_\infty + ||\overline{Q}_k - \mathcal{T}\underline{Q}_k||_\infty \\
&= ||\underline{Q}_k - \overline{Q}_k||_\infty + \max_{s,a}|\overline{Q}_k(s,a) - \mathcal{T}\underline{Q}_k(s,a)| \\
&\leq ||\underline{Q}_k - \overline{Q}_k||_\infty + \gamma\rho_{k-1} \\
&= \frac{1}{k}||Q_k - Q_0||_\infty + \gamma\rho_{k-1} \\
&\leq \frac{1}{k}(1/(1-\gamma)||r||_\infty + 2||Q_0||_\infty) + \gamma\rho_{k-1}
\end{aligned}$$

□

It remains to show that such a converging sequence of approximate fixed points converges to Q^*, the fixed point of $\mathcal{T}$.

Lemma 4. *Let* $Q_0, Q_1, \ldots$ *be a sequence such that* $\lim_{k\to\infty}||Q_k - \mathcal{T}Q_k||_\infty = 0$. *Then* $\lim_{k\to\infty} Q_k = Q^*$.

Proof.

$$\begin{aligned}
||Q_k - Q^*||_\infty &\leq ||Q_k - \mathcal{T}Q_k||_\infty + ||\mathcal{T}Q_k - Q^*||_\infty \\
&= ||Q_k - \mathcal{T}Q_k||_\infty + ||\mathcal{T}Q_k - \mathcal{T}Q^*||_\infty \\
&\leq ||Q_k - \mathcal{T}Q_k||_\infty + \gamma||Q_k - Q^*||_\infty \\
&= \frac{1}{1-\gamma}||Q_k - \mathcal{T}Q_k||_\infty
\end{aligned}$$

Thus, by assumption, $\limsup_{k\to\infty}||Q_k - Q^*||_\infty \leq 0$. Since $||Q_k - Q^*||_\infty \geq 0$, $\liminf_{k\to\infty}||Q_k - Q^*||_\infty \geq 0$. Thus $\lim_{k\to\infty}||Q_k - Q^*||_\infty = 0$ and the result follows. □

Combining Lemmas 3 and 4 shows the convergence of LONR learning.

Theorem 1. $\lim_{k\to\infty}\underline{Q}_k = Q^*$.

Discussion

We have proposed a new learning algorithm for MDPs, local no-regret learning (LONR), and shown that it has the same convergence guarantees as Q-learning when complete updates are performed on all states simultaneously. However, Q-learning also converges when updates are done via sampling, and this is important in practice for problems of more than moderate size. Also, for standard MDPs, Q-learning already suffices and LONR provides no obvious advantage. Intuitively it could help with multiple agents and partial observability, but our existing analysis does not address these extensions. In the remainder of the paper, we discuss each of these issues and directions towards addressing them.

Like LONR, CFR was originally developed to update all states simultaneously. However, with some clever techniques to reduce the computational and space overhead of this approach, it has shown success in solving games of moderate size (Bowling et al. 2015). However, a line of work has also shown that CFR will also converge when sampling trajectories (Lanctot et al. 2009; Gibson et al. 2012; Johanson et al. 2012), and it is plausible that this approach

could be applied to LONR as well to provide a sampling version. Another approach, which has a strong intuitive appeal, is to observe that LONR uses an arbitrary no-absolute regret learning algorithm for the full information case, where the algorithm can observe the reward for each action. No-regret algorithms are also studied for the "bandit" case, where only the reward for the chosen action is observed. This matches the information structure of sampling, with the additional complication that we also only get samples of the reward and next state rather than its expectation and the full distribution respectively. Thus, a natural conjecture is that simply using a bandit algorithm as our no-absolute-regret learning algorithm would lead to convergence with sampling. As this would require a substantially different analysis to our current approach, we leave this as an open problem.

In addition to sampling, current algorithms such as DQN (Mnih et al. 2015) rely on approximation. While there has been some exploration of function approximation in the context of CFR (Waugh et al. 2015; Moravčík et al. 2017; Jin, Levine, and Keutzer 2017), more work is needed, and the right way to combine with a more general framework like LONR, remains an open question.

Multiple agents and partial observability introduce similar issues in that P and r are no longer stationary and in general we have a P_k and R_k for each round k. This causes problems with the proof of Lemma 1. In particular, note that in the third step of the proof we interchange the sum over rounds and P and r to be able to apply the no-regret property in the fourth line. Without the ability to do this, we end up needing to apply the no-regret property to a weighted sum, and in general the weighting may differ from state to state, which makes generalizing the proof non-trivial. However, there is both theoretical and empirical cause for hope. On the theoretical side, CFR provides an example of such a guarantee, although there the special structure that the state contains the full history of play (and thus the states form a tree structure) is exploited. On the empirical side, we have the successes of both CFR (Brown and Sandholm 2017; Moravčík et al. 2017) and ARM (Jin, Levine, and Keutzer 2017), in particular despite the lack of any theoretical guarantees for the latter. It is possible that this approach requires at least some restriction on the domain, and in the appendix we explore a special case (which captures among other things normal form games) where a variant of Lemma 1 can be proved. Proving more general versions remains an open problem.

A more narrow technical issue is that our argument relied on a slightly stronger version of the usual no-regret property, which we termed no-absolute-regret. Is this necessary or an artifact of our proof technique? Relatedly, while several algorithms are known which satisfy this stronger property, we are not sure whether regret matching (the algorithm used by CFR) does or not.

References

Bellemare, M. G.; Ostrovski, G.; Guez, A.; Thomas, P. S.; and Munos, R. 2016. Increasing the action gap: New operators for reinforcement learning. In *AAAI*, 1476–1483.

Bertsekas, D. P., and Tsitsiklis, J. 1996. *Neuro-Dynamic Programming*. Athena Scientifit.

Bowling, M.; Burch, N.; Johanson, M.; and Tammelin, O. 2015. Heads-up limit hold'em poker is solved. *Science* 347(6218):145–149.

Brown, N., and Sandholm, T. 2017. Libratus: the superhuman ai for no-limit poker. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Daumé III, H.; Langford, J.; and Sharaf, A. 2018. Residual loss prediction: Reinforcement learning with no incremental feedback.

Even-Dar, E.; Kakade, S. M.; and Mansour, Y. 2005. Experts in a markov decision process. In *Advances in neural information processing systems*, 401–408.

Even-Dar, E.; Kakade, S. M.; and Mansour, Y. 2009. Online markov decision processes. *Mathematics of Operations Research* 34(3):726–736.

Gibson, R. G.; Lanctot, M.; Burch, N.; Szafron, D.; and Bowling, M. 2012. Generalized sampling and variance in counterfactual regret minimization. In *AAAI*.

Gofer, E., and Mansour, Y. 2016. Lower bounds on individual sequence regret. *Machine Learning* 103(1):1–26.

Gondek, D.; Greenwald, A.; and Hall, K. 2004. Qnr-learning in markov games.

Greenwald, A., and Jafari, A. 2003. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Learning Theory and Kernel Machines*. Springer. 2–12.

Greenwald, A.; Hall, K.; and Serrano, R. 2003. Correlated q-learning. In *ICML*, volume 3, 242–249.

Hu, J., and Wellman, M. P. 2003. Nash q-learning for general-sum stochastic games. *Journal of machine learning research* 4(Nov):1039–1069.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(Apr):1563–1600.

Jin, P. H.; Levine, S.; and Keutzer, K. 2017. Regret minimization for partially observable deep reinforcement learning. *arXiv preprint arXiv:1710.11424*.

Johanson, M.; Bard, N.; Lanctot, M.; Gibson, R.; and Bowling, M. 2012. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 837–846. International Foundation for Autonomous Agents and Multiagent Systems.

Kaufmann, E., and Koolen, W. M. 2017. Monte-carlo tree search by best arm identification. In *Advances in Neural Information Processing Systems*, 4904–4913.

Lanctot, M.; Waugh, K.; Zinkevich, M.; and Bowling, M. 2009. Monte carlo sampling for regret minimization in extensive games. In *Advances in neural information processing systems*, 1078–1086.

Ma, Y.; Zhang, H.; and Sugiyama, M. 2015. Online markov

decision processes with policy iteration. *arXiv preprint arXiv:1510.04454*.

Mannor, S., and Shimkin, N. 2003. The empirical bayes envelope and regret minimization in competitive markov decision processes. *Mathematics of Operations Research* 28(2):327–345.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Montgomery, W.; Ajay, A.; Finn, C.; Abbeel, P.; and Levine, S. 2017. Reset-free guided policy search: Efficient deep reinforcement learning with stochastic initial states. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, 3373–3380. IEEE.

Moravčík, M.; Schmid, M.; Burch, N.; Lisỳ, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356(6337):508–513.

Osband, I., and Van Roy, B. 2016. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 627–635.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484–489.

Van Seijen, H.; Van Hasselt, H.; Whiteson, S.; and Wiering, M. 2009. A theoretical and empirical analysis of expected sarsa. In *Adaptive Dynamic Programming and Reinforcement Learning, 2009. ADPRL'09. IEEE Symposium on*, 177–184. IEEE.

Waugh, K.; Morrill, D.; Bagnell, J. A.; and Bowling, M. 2015. Solving games with functional regret estimation. In *AAAI*, volume 15, 2138–2144.

Yu, J. Y.; Mannor, S.; and Shimkin, N. 2009. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research* 34(3):737–757.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2008. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, 1729–1736.

Beyond MDPs

If we move beyond MDPs, P and r are no longer stationary and in general we have a P_k and R_k. This causes problems with the proof of Lemma 1. Recall the initial part of that proof, updated to this more general setting:

$$\begin{aligned}\overline{Q}_k(s,a) &= \frac{1}{k}\sum_{t=1}^{k} Q_t(s,a) \\ &= \frac{1}{k}\sum_{t=0}^{k-1} r_t(s,a) + \gamma \mathbb{E}_{P_t,\pi_t}[Q_t(s',a')]\end{aligned}$$

In the original proof, we pulled the expectation over P outside the sum, but now we cannot. In particular, writing the expecation more explicitly gives

$$\frac{1}{k}\sum_{t=0}^{k-1} r_t(s,a) + \gamma \sum_{s'\in\mathcal{S}} P_t(s' \mid s,a)\mathbb{E}_{\pi_t}[Q_t(s',a')] \quad (4)$$

We can still reverse the order of the sums, but the weighting terms now depend on t so they cannot be moved outside. More problematically, they also depend on s and a, so it is not immediately clear how to generalize our results.

For intuition about the sort of problems that could arise, consider a state s' where there are two actions. At odd k, $r_k(s',a_1) = 1$ and $r_k(s',a_2) = 0$ and vice versa at even k. It is a valid no-regret strategy to randomize uniformly over the actions, but if the P_k are such that you only arrive in s' from s at odd k, then this gives an incorrect estimate.

In the remainder of this section, we analyze a special case where we can prove a variant of Lemma 1.

Time-invariant P

If P does not change with k, but r does, we can still prove a version of Lemma 1. With a single state, this captures learning in normal-form games, where no-regret learning is indeed known to work. This assumption is also common in the literature on "online MDPs" (Even-Dar, Kakade, and Mansour 2009; Mannor and Shimkin 2003; Yu, Mannor, and Shimkin 2009; Ma, Zhang, and Sugiyama 2015) In this setting, a version of Lemma 1 can be proved, but now rather than having a constant operator $\mathcal{T}$ it now changes over time as

$$\mathcal{T}_k Q(s,a) = \underline{r}_k(s,a) + \gamma \mathbb{E}_P[\max_i Q(s',a_i)]. \quad (5)$$

Lemma 5.

$$-\gamma\rho_{k-1} + \mathcal{T}_k \underline{Q}_k(s,a) \le \overline{Q}_k(s,a) \le \gamma\rho_{k-1} + \mathcal{T}_k \underline{Q}_k(s,a). \quad (6)$$

Proof.

$$
\begin{aligned}
\overline{Q}_k(s,a) &= \frac{1}{k}\sum_{t=1}^{k} Q_t(s,a) \\
&= \frac{1}{k}\sum_{t=0}^{k-1} r_t(s,a) + \gamma \mathbb{E}_{P,\pi_t}[Q_t(s',a')] \\
&= \frac{1}{k}\sum_{t=0}^{k-1} r_t(s,a) + \gamma \mathbb{E}_{P}[\frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}_{\pi_t}[Q_t(s',a')]] \\
&\geq \frac{1}{k}\sum_{t=0}^{k-1} r_t(s,a) + \gamma \mathbb{E}_{P}[\max_i \frac{1}{k}\sum_{t=0}^{k-1} Q_t(s',a_i) - \rho_{k-1}] \\
&= -\gamma\rho_{k-1} + \frac{1}{k}\sum_{t=0}^{k-1} r_t(s,a) + \gamma \mathbb{E}_{P}[\max_i \frac{1}{k}\sum_{t=0}^{k-1} Q_t(s',a_i)] \\
&= -\gamma\rho_{k-1} + \underline{\mathrm{r}}_k(s,a) + \gamma \mathbb{E}_{P}[\max_i \underline{Q}_k(s',a_i)] \\
&= -\gamma\rho_{k-1} + \mathcal{T}_k \underline{Q}_k(s,a)
\end{aligned}
$$

As before, the key step is applying the no-regret property to obtain the inequality and we apply the same argument with the no-absolute-regret property to obtain the reverse inequality. □