Extreme State Aggregation beyond MDPs

Marcus Hutter

Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

Abstract

I consider a reinforcement learning setup without any (esp. MDP) assumptions on the environment. State aggregation and more generally feature reinforcement learning is concerned with mapping histories/raw-states to aggregated states. The idea behind both is that the resulting reduced process (approximately) forms a small stationary finite-state MDP, which can then be efficiently solved or learnt. I considerably generalize existing aggregation results by showing that even if the reduced process is not an MDP, the (q)value functions and (optimal) policies of an associated MDP with same state-space size solve the original problem, as long as the solution can approximately be represented as a function of the reduced states. This implies an upper bound on the required state space size that holds uniformly for all RL problems. It may also explain why RL algorithms designed for MDPs sometimes perform well beyond MDPs.

Overview in 1 Slide

- Setup: Reinforcement Learning (RL) without any (esp. MDP) assumptions on the environment. Is very hard problem. Approaches:
- State aggregation: Partitions (raw) states into fewer aggr. states.
- Feature Reinforcement Learning: maps/reduces histories to states.
- So far: Resulting process needed to (approximately) form a small stationary finite-state MDP, which can then be efficiently solved.
- New: Even if the reduced process is not an MDP, there is an *associated MDP* of same size whose optimal value&policy approximately solve the original problem.
- Only condition: Solution can still be approximately represented.
- Implications: Uniform upper bound on the required state space size for *all* RL problems.
- Explains why RL algorithms designed for MDPs sometimes perform well beyond MDPs.

Contents

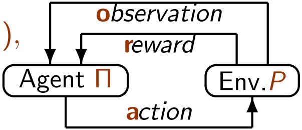
- Feature Markov Decision Processes
- Approximate Aggregation for General Processes
- Extreme Aggregation
- Reinforcement Learning
- Feature Reinforcement Learning
- Discussion

Agent-Environment Setup [Hut09]

- Agent Π interacts with an Environment P:
  actions $a \in \mathcal{A}$, observations $o \in \mathcal{O}$, real-valued rewards $r \in \mathcal{R} \subseteq [0;1]$

Env. $P : \mathcal{H} \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$, $\quad P(o_{t+1} r_{t+1} | h_t a_t)$,

Agent $\Pi : \mathcal{H} \to \mathcal{A}$, $\quad a_t = \Pi(h_t)$,



- (Π, P) generate history $h \in \mathcal{H} := (\mathcal{O} \times \mathcal{R} \times \mathcal{A})^* \times \mathcal{O} \times \mathcal{R}$:
  $h_t \;:=\; o_1 r_1 a_1 ... o_{t-1} r_{t-1} a_{t-1} o_t r_t \;\in\; \mathcal{H}_t := (\mathcal{O} \times \mathcal{R} \times \mathcal{A})^{t-1} \times \mathcal{O} \times \mathcal{R}$

- $\mathcal{O}$ and $\mathcal{R}$ and $\mathcal{A}$ assumed to be finite.

- Agent's objective is to maximize its long-term reward.

- We make no (stationarity or Markov or other) assumption on environment P.

(Optimal) Value Functions, Policies, and History Bellman Equations

Performance of a policy Π is measured in terms of the expected γ-discounted reward, called (Q)-Value of Π at history h_t (and action a_t)

$$\begin{aligned} V^\Pi(h_t) &:= \mathbb{E}^\Pi[R_{t+1}|h_t] \\ Q^\Pi(h_t,a_t) &:= \mathbb{E}^\Pi[R_{t+1}|h_t a_t] \\ R_t &:= \textstyle\sum_{\tau=t}^\infty \gamma^{\tau-t} r_\tau \end{aligned}$$

The Optimal Policy and (Q)-Value functions are

$$\begin{aligned} V^*(h_t) &:= \max_\Pi V^\Pi(h_t) \\ Q^*(h_t,a_t) &:= \max_\Pi Q^\Pi(h_t,a_t) \\ \Pi^* &:\in \arg\max_\Pi V^\Pi(\epsilon) \end{aligned}$$

From Histories to States (ϕ)

- Space of histories is huge and unwieldy and no history ever repeats.
  $\Longrightarrow$ Problem: Prevents naive learning based on frequencies.

- Solution: Aggregate similar histories: Feature map $\phi : \mathcal{H} \to \mathcal{S}$ reduces histories $h_t \in \mathcal{H}$ to states $s_t := \phi(h_t) \in \mathcal{S}$.

- The probability of successor states and rewards can be obtained by marginalization: $P_\phi(s_{t+1} r_{t+1} | h_t a_t) := \sum_{\tilde{o}_{t+1} : \phi(h_t a_t \tilde{o}_{t+1} r_{t+1}) = s_{t+1}} P(\tilde{o}_{t+1} r_{t+1} | h_t a_t)$

- We **neither** assume P_ϕ to be MDP **nor** to be stationary.

Classical State Aggregation

- Assumes P is MDP in observations: $P(o'r'|ha) = P(o'r'|oa)$
- Aggregates $s_t = \phi(o_t)$ via equivalent partitioning: $\{\phi^{-1}(s) : s \in \mathcal{S}\}$
- s_t is supposed to summarize all relevant information from obs. o_t.
- Formally: Assumes P_ϕ is (approximately) a stationary MDP (bisimulation condition [GDG03, FPP04]):

$$P_\phi \in \text{MDP} \;\; :\Leftrightarrow \;\; \exists p : P_\phi(s_{t+1}r_{t+1}|\tilde{h}_t a_t) = p(s_{t+1}r_{t+1}|s_t a_t) \;\; \forall \phi(\tilde{h}_t) = s_t$$

This is precisely the condition we lift.

Markov Decision Processes (MDP)

- Upper-case letters V, Q, Π for the general process P.
  Lower-case letters v, q, π for stationary MDP p.

- Consider a stationary finite-state MDP $p : \mathcal{S}\times\mathcal{A} \rightsquigarrow \mathcal{S}\times\mathcal{R}$, and its (stationary deterministic) optimal policy $\pi^* : \mathcal{S}\to\mathcal{A}$.

- MDP Bellman optimality equations:

$$\begin{aligned} q^*(s,a) &= \sum_{s'r'} p(s'r'|sa)[r'+\gamma v^*(s')] \\ v^*(s) &= \max_a q^*(s,a) \\ \pi^*(s) &\in \arg\max_a q^*(s,a) \end{aligned}$$

- If P reduces via ϕ to an MDP $p = P_\phi$, then the solution of these equations, yields optimal (Q)-Values and optimal Policy of the original process P. But in general $p \neq P_\phi$!

Dispersion Probability B

- $B:\mathcal{S}\times\mathcal{A}\rightsquigarrow\mathcal{H}$ may be viewed as a (weird) stochastic inverse of ϕ that assigns non-zero probability (only) to $h\in\phi^{-1}(s)$ of any/mixed length:

$$B(h|sa)\geq 0 \quad\text{and}\quad \sum_{h\in\mathcal{H}}B(h|sa)=\sum_{h:\phi(h)=s}B(h|sa)=1 \quad \forall s,a \qquad (1)$$

- Definition: $$p(s'r'|sa) \;:=\; \sum_{h\in\mathcal{H}}P_\phi(s'r'|ha)B(h|sa) \qquad (2)$$

- p is a stationary MDP for any B satisfying (1) and any ϕ and P.

- Easy to see: $P_\phi\in\text{MDP} \quad\Longleftrightarrow\quad p=P_\phi$ (any B)

- In general p is not the state distribution induced by P (and Π), which in general is non-Markov.

Relating P and p via B

Key relation between P and p via B used later to relate original history with reduced state Bellman equations.

Lemma (**BPp**)

For any function $f : \mathcal{S} \times \mathcal{R} \to \mathbb{R}$ *and* p *defined in (2) in terms of* P, *and* $s' := \phi(h')$ *and* $h' := hao'r'$ *it holds*

$$\sum_{h\in\mathcal{H}} B(h|sa) \sum_{o'r'} P(o'r'|ha) f(\underset{\substack{\uparrow \\ \textit{depends on } o'r'}}{s'}, r') \;=\; \sum_{s'r'} p(s'r'|sa) f(s', r')$$

Relating $\mathbf{v}-\mathbf{V}$ and $\mathbf{q}-\mathbf{Q}$

Lemma ($|\mathbf{v}-\mathbf{V}| \leq \max_a |\mathbf{q}-\mathbf{Q}|$ and $|\mathbf{q}-\langle\mathbf{Q}\rangle_B| \leq \gamma|\mathbf{v}-\mathbf{V}|$)

(i) $|v^*(s)-V^*(h)| \;\leq\; \max_a |q^*(s,a)-Q^*(h,a)| \quad \forall s,h,a$

(ii) *For any P, ϕ, and B, define p via (2). Then*
$|q^*(s,a)-\langle Q^*(h,a)\rangle_B| \;\leq\; \gamma|v^*(s)-V^*(h)| \quad \forall s=\phi(h)\;\forall a,$
where $\langle f(h,a)\rangle_B := \sum_{\tilde h\in\mathcal{H}} B(\tilde h|sa)f(\tilde h,a)$ *with* $s:=\phi(h)$

- (i) trivially bounds $v-V$ differences in terms of $q-Q$ differences.
- (ii) non-trivially shows that a reverse holds in expectation.
- $\langle f(h,a)\rangle_B$ takes a B-average over all $\tilde h$ that ϕ maps to same state as h
- Function $f(h)$ is called ϕ-uniform iff $f(h)=f(\tilde h)$ for all $\phi(h)=\phi(\tilde h)$.
- Expectation can (only) be dropped if Q^* is ϕ-uniform.

Main Result

Theorem ($\phi\mathbf{Q}*$)

For any P, ϕ, and B, define p via (2).
Assume $|Q^*(h,a)-Q^*(\tilde h,a)|\leq\varepsilon$ for all $\phi(h)=\phi(\tilde h)$ and all a.
Then for all a and h and $s=\phi(h)$ it holds:

$$(i)\quad |Q^*(h,a)-q^*(s,a)|\leq\tfrac{\varepsilon}{1-\gamma}\quad\textit{and}\quad |V^*(h)-v^*(s)|\leq\tfrac{\varepsilon}{1-\gamma},$$

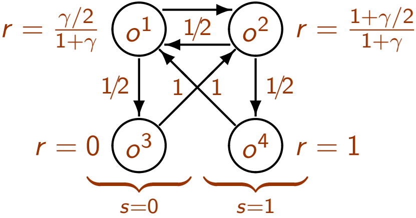
$$(ii)\quad 0\ \leq\ V^*(h)-V^{\tilde\Pi}(h)\ \leq\ \tfrac{2\varepsilon}{(1-\gamma)^2},\quad\textit{where}\quad \tilde\Pi(h):=\pi^*(s)$$

$$(iii)\quad \textit{If } \varepsilon=0 \textit{ then } \Pi^*(h)=\pi^*(s)$$

- Meaning: We can aggregate histories as much as we wish, as long as the optimal value function and policy are still approximately representable as functions of aggregated states.
- Whether the reduced process P_ϕ is Markov is immaterial.
  We can use surrogate MDP p to find an ε-optimal policy for P.
- Similar results hold for V^Π, Q^Π, V^*, but some questions are open.

Simple Example

- A simple example of a P and ϕ that satisfy the conditions Theorem 3,
- but violate the bisimulation condition [GDG03]
- and indeed have large bisimulation distance [FPP04].


$r = \frac{\gamma/2}{1+\gamma}$
o^1
1/2
o^2
$r = \frac{1+\gamma/2}{1+\gamma}$
1/2
1
1
1/2
$r = 0$
o^3
o^4
$r = 1$
$s{=}0$
$s{=}1$

- (Q)Value function $V(o_t) := V^*(h_t) = Q^*(h_t, a_t)$ is ϕ-uniform: $V(o^1) = V(o^3) = \frac{\gamma}{1-\gamma^2}$ and $V(o^2) = V(o^4) = \frac{1}{1-\gamma^2}$.
- Theorem 3 can be applied to aggregate the four raw states $\mathcal{O} = \{o^1, ..., o^4\}$ into two states $\mathcal{S} = \{0, 1\}$.

Extreme Aggregation

- Thm.3 allows to represent **any** Process as a small finite-state MDP.
- Consider ϕ that maps each history to the vector-over-actions of optimal Q-values $Q^*(h,\cdot)$ discretized to some finite ε-grid:

$$\phi(h) \;:=\; \big(\lfloor Q^*(h,a)/\varepsilon\rfloor\big)_{a\in\mathcal{A}} \;\in\; \{0,1,...,\lfloor\tfrac{1}{\varepsilon(1-\gamma)}\rfloor\}^{\mathcal{A}} \;=:\; \mathcal{S} \qquad (3)$$

  I.e. all histories with ε-close Q^*-values are mapped to the same state:
- Now find π^* of MDP p of size $|\mathcal{S}|$ and define $\tilde{\Pi}(h) := \pi^*(\phi(h))$.
- By Thm.3ii, $\tilde{\Pi}$ is an $2\varepsilon/(1-\gamma)^2$-optimal policy of original process P.

Theorem (Extreme ϕ)

For every process P, reduction ϕ (3) and MDP p (2) has optimal policy π^*, which is an ε-optimal policy $\tilde{\Pi}(h) := \pi^*(\phi(h))$ for P. The size of the MDP is bounded (uniformly for **any** *P) by $|\mathcal{S}| \leq \left(\frac{3}{\varepsilon(1-\gamma)^3}\right)^{|\mathcal{A}|}$*

Discussion of Extreme Aggregation

- We do not know Q^* in advance, so what are these results good for?
- Start with a sufficiently rich class of maps Φ that contains at least one ϕ approximately representing $Q^*(h,\cdot)$,
- Have a learning algorithm that favors such ϕ,
- Then Theorem 3 tells us that we do not need to worry about whether P_ϕ is MDP or not; we “simply” use/learn MDP p instead.
- Theorem 4 tells us that this allows for extreme aggregation way beyond MDPs.
- Conjecture: If $\phi(h) := \big(\lfloor V^*(h)/\varepsilon \rfloor, \Pi^*(h)\big) \in \{0,1,...,\lfloor\frac{1}{\varepsilon(1-\gamma)}\rfloor\} \times \mathcal{A} =: \mathcal{S}$ then $|V^{\tilde\Pi(h)} - V^*(h)| = O(\varepsilon)$ hence $|\mathcal{S}| = O(|\mathcal{A}|/\varepsilon)$ i.e. $|\mathcal{S}|$ is only linear in $|\mathcal{A}|$, not exponential as in Thm.4.

Choice of B

- Let $\Pi_B : \mathcal{H} \rightsquigarrow \mathcal{A}$ be a general behavior policy of our RL Agent.
- The (Π_B, P)-interaction generates joint probability, say $P_B(h_t a_t)$.
- Subscripts B and ϕ indicate dependence on Π_B and/or ϕ.
- By marginalization and conditioning we get $P_{\phi B}(h_t|s_t a_t)$ in the usual way, and similar for other arguments.
- Introduce weights $w_t : \mathcal{S} \times \mathcal{A} \rightsquigarrow [0;1]$ and define

$$B(h_t|sa) := w_t(sa)P_{\phi B}(h_t|s_t=s, a_t=a) \ \forall t, \text{where } \sum_{t=1}^{\infty} w_t(sa)=1 \ \forall s, a$$

- B satisfies (1) and leads to $p(s'r'|sa) =$

$$= \sum_{t=1}^{\infty} w_t(sa) \sum_{h_t \in \mathcal{H}_t} P_\phi(s_{t+1}=s', r_{t+1}=r'|h_t, a_t=a) P_{\phi B}(h_t|s_t=s, a_t=a)$$

$$= \sum_{t=1}^{\infty} w_t(sa) P^t_{\phi B}(s'r'|sa).$$

$P^t_{\phi B}$ cannot be estimated, but ...

Estimation of p

- Choose $w_t(sa) := \frac{P^t_{\phi B}(sa)}{\sum_{t=1}^n P^t_{\phi B}(sa)}$ for $t \leq n$ and 0 for $t > n$

$$\Longrightarrow p(s'r'|sa) = \frac{\frac{1}{n}\sum_{t=1}^n P^t_{\phi B}(sas'r')}{\frac{1}{n}\sum_{t=1}^n P^t_{\phi B}(sa)}$$

- Under weak conditions this can be estimated as follows:

Count number of times action a is taken in state s: $\quad n(sa) := \sum_{t=1}^n X_t = \sum_{t=1}^n [[s_t = s \wedge a_t = a]]$

- $\mathbb{E}[X_t] = P(X_t = 1) = P^t_{\phi B}(sa)$

- Similarly: $n(sas'r') := \sum_{t=1}^n Y_t$, where $Y_t := [[s_t a_t s_{t+1} r_{t+1} = asa'r']]$

Theorem (p-estimation)

$$\frac{n(sas'r')}{n(sa)} - p(s'r'|sa) \xrightarrow{n(sa)\to\infty} 0 \quad \textit{a.s. under weak conditions.}$$

For example, convergence holds if Y_t are stationary ergodic processes.

Discussion of p-Estimation

- Limit $n()/n() \to p()$ shows that standard frequency estimation for p will converge to the true p under weak conditions.
- If P_ϕ is MDP, samples are conditionally i.i.d. and the 'weak conditions' are satisfied.
- But Laws of Large Numbers hold way beyond the i.i.d. case [FK01].
- Model-free learning possible too: Condition $n()/n() \to p()$ should be sufficient for Q-learning to converge to Q^*.
- Q-learning and other RL algorithms designed for MDPs have been observed to often (but not always) perform well even if applied to non-MDP domains. Our results appear to explain why.

Feature Reinforcement Learning

- The idea of FRL is to **learn** ϕ [Hut09].
- FRL starts with a class of maps Φ, compares different $\phi \in \Phi$, and selects the most appropriate one given the experience h_t so far.
- Several criteria based on how well ϕ reduces P to an MDP have been devised.
- Theorems 3 shows that demanding P_ϕ to be approximately MDP is overly restrictive.
- Theorem 4 suggests that if we relax this condition, much more substantial aggregation is possible, provided Φ is rich enough.

Search for Exact ϕ based on Infinite Sample Size

- We call a reduction $\phi : \mathcal{H} \to \mathcal{S}_\phi$ exact iff $Q^*(h,a) = q^*_\phi(s,a)$ and $\Pi^*(h) = \pi^*_\phi(s)$ for all $s = \phi(h)$ and a.
- Even for $n = \infty$, P hence Q^* needed for Π^* is (usually) not estimable (from h_∞).
- On the other hand, for each $\phi \in \Phi$, $p = p_\phi$ can be determined (exactly) (under weak conditions).
- From p_ϕ we can determine q^*_ϕ and π^*_ϕ via (1).
- The solution always satisfies the reduced Bellman equations exactly, even for very bad reductions, e.g. single state $\phi(h) \equiv 0 \, \forall h$.
- So the reduced problem is not sufficient to judge the quality of ϕ.

Search for Exact ϕ based on Infinite Sample Size

- Coarsening and refining reductions ϕ:
  Let us now coarsen ϕ: Consider $\chi : \mathcal{S}_\phi \to \mathcal{S}_\psi$ and $\psi : \mathcal{H} \to \mathcal{S}_\psi$ such that $\psi(h) = \chi(\phi(h))$.
  Example: Splitting/marging nodes in tree representation of states.

- Partially order reductions in Φ:
  $\psi \prec \phi :\Leftrightarrow q^*_\phi$ and π^*_ϕ are constant on all $s_\phi \in \chi^{-1}(s_\psi)$ for all s_ψ and a

- Enriching the order: $\psi \prec_\times \psi'$:iff $\psi \prec \phi \prec \psi'$ or $\psi \prec \psi'$, $\phi := (\psi, \psi')$

- Search for ϕ: Assume Φ contains at least one exact reduction and is closed under arbitrary coarsening $\implies$ $\exists$ unique $\prec_\times$-minimizer ϕ_0.

- Theorem 3 justifies $\prec_\times$-minimization
  based on (q^*_ϕ, π^*_ϕ) that ignores the (non)Markov structure of P_ϕ.

Search for Approximate ϕ based on Finite Sample Size

The principle approach in the previous paragraph is sound,
but needs to be generalized in various ways before it can be used:

- Approximate equality: $\hat{q}^*_\phi \approx \hat{q}^*_\psi$
- Finite sample size: e.g. Kolmogorov-Smirnov test
- Exploration: optimism
- Regularization: penalizing complex ϕ
- Efficient search: heuristic rather than exhaustive search for ϕ_0

All but the last point raised above have or should have general solutions (see next slide).

Utilizing Existing Algorithms

- The BLB algorithm family [Ngu13] solves most of the problems above and can be used/adapted for our purpose.
- BLB algorithm and analysis relies on UCRL2 [JOA10], an exploration algorithm for finite-state MDPs.
- Replacing P_ϕ by p in the proof of BLB and UCRL2 should work.
- UCRL2 analysis exploits that s', r' conditioned on s, a are i.i.d., which is true no longer true for $P_\phi \notin \mathrm{MDP}$.
- Hoeffding's inequality for i.i.d. needs to be replaced by comparable bounds with weaker conditions, e.g. Azuma's inequality for martingales.
- Problem: BLB considers average reward and regret, while our theorems are for discounted reward.
- ToDo: Derive PAC version of BLB for discounted reward, e.g. by combining MERL [LHS13] with UCRLγ [LH12].

Summary

- Our results show that RL algorithms for finite-state MDPs can be utilized even for problems P that have arbitrary history dependence and history-to-state reductions/aggregations ϕ that induce P_ϕ that are also neither stationary nor MDP.

- The only condition to be placed on the reduction is that the quantities of interest, (Q)Values and (optimal) Policies, can approximately be represented.

- This considerably generalizes previous work on Feature Reinforcement Learning and MDP state aggregation and allows for extreme state aggregations beyond MDPs.

- The obtained results may also explain why RL algorithms designed for MDPs sometimes perform well beyond MDPs.

Outlook

- Weaken condition on Q^* to V^* in Theorem 3ii.
- Develop algorithm learning ϕ beyond MDPs that comes with PAC guarantees (e.g. MERL+UCRLγ).
- All bounds contain $\frac{1}{1-\gamma}$ to some power. Are they tight?
- Generalize exact to approximate ϕ-uniformity conditions for given Π.
- Use new theorems and/or proof ideas to extend existing convergence theorems for RL algorithms such as Q-learning and others from MDPs to beyond MDPs.

References

M. Hutter.
Extreme state aggregation beyond MDPs.
In Proc. ALT'14, volume 8776 of LNAI, pages 185–199, 2014.

M. Hutter.
Feature reinforcement learning: Part I: Unstructured MDPs.
Journal of Artificial General Intelligence, 1:3–24, 2009.

T. Lattimore and M. Hutter.
PAC bounds for discounted MDPs.
In Proc. ALT'12, volume 7568 of LNAI, pages 320–334, 2012.

T. Lattimore, M. Hutter, and P. Sunehag.
The sample-complexity of general reinforcement learning.
Journal of Machine Learning Research, W&CP: ICML, 28(3):28–36, 2013.

P. Nguyen.
Feature Reinforcement Learning Agents.
PhD thesis, RSCS, Australian National University, 2013.