Conditions on Features for Temporal Difference-like Methods to Converge

Marcus Hutter, Samuel Yang-Zhao, Sultan Majeed [1]

August 4, 2019

[1] College of Engineering and Computer Science, Australian National University

Overview

- Investigate model-free, 'off-policy', convergence to the correct solution for natural algorithms.
- Natural algorithms are RL methods with linear function approximation that take a projection on the Bellman equation.
- E.g. $TD(\lambda)$, Q-learning, $GTD(\lambda)$.

Overview continued...

- Who might care about our work? Theoretical RL researchers.
- What's new in our approach?
  - We provide a complete, theoretical characterization of convergence based on the choice of features.
  - State aggregation is proven to be a feature construction choice that will always converge.
  - A condition on finding convergent algorithms beyond state aggregation is provided.

Overview continued...

Advantages:

- Analysis is model and policy agnostic.
- Results hold for a large class of RL algorithms (the natural algorithms).

Disadvantages:

- Whilst extensive, natural algorithms do not cover all RL algorithms with linear function approximation.

Problem

The convergence proofs of many reinforcement learning algorithms with linear function approximation assume uniqueness of solution.

Example (Convergence of *GTD*2 (Sutton et al., 2009))

Theorem 1 (Convergence of GTD2). *Consider the GTD2 iterations (8) and (9) with step-size sequences α_k and β_k satisfying $\beta_k = \eta\alpha_k$, $\eta > 0$, $\alpha_k, \beta_k \in (0,1]$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Further assume that (ϕ_k, r_k, ϕ_k') is an i.i.d. sequence with uniformly bounded second moments. Let $A = \mathbb{E}\left[\phi_k(\phi_k - \gamma\phi_k')^\top\right]$, $b = \mathbb{E}[r_k\phi_k]$, and $C = \mathbb{E}\left[\phi_k\phi_k^\top\right]$. Assume that A and C are non-singular. Then the parameter vector θ_k converges with probability one to the TD fixpoint (4).*

- The matrix quantities A and C are functions of the chosen features.
- Making C invertible is easy, A is much harder to guarantee.

We give a more general characterization.

Problem

Do there exist conditions on features that can characterize uniqueness, and hence convergence, for reinforcement learning algorithms?

Relevant Research

Relevant counter-examples and the theory of linear value function approximation have been developed:

- $TD(0)$ diverges under off-policy learning even if value function can be represented exactly (Tsitsiklis and Van Roy, 1997).
- Linear value function approximation unified in an oblique projection framework (Scherrer, 2010).
- Counter-example showing non-uniqueness of Bellman error solutions (Sutton and Barto, 2018).

A Non-uniqueness Example (Sutton and Barto, 2018)

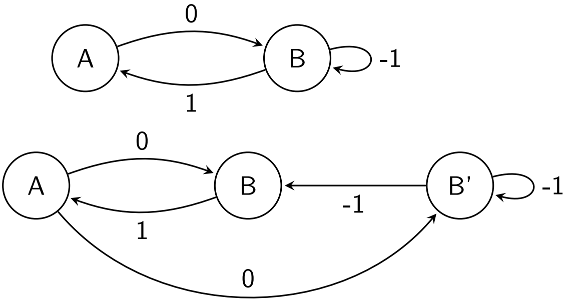


Figure: Counter-example showing Bellman error methods suffer from non-uniqueness.

- Use a two component parameter vector to represent the value function in both MDPs.
- The observable feature-reward sequence is the same for both MDPs $\Longrightarrow$ the observed data distribution is identical.

A Non-uniqueness Example (Sutton and Barto, 2018)

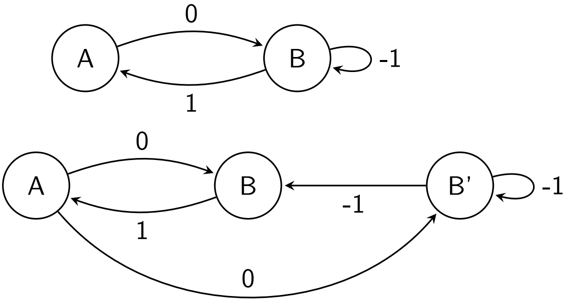


Figure: Counter-example showing Bellman error methods suffer from non-uniqueness.

- For a parameter value of 0, the Bellman errors in each MDP differ: $\overline{BE}_1 = 0$ and $\overline{BE}_2 = \frac{2}{3}$.
- Conclusion: Converging to the minimum Bellman error may lead to the wrong parameter!

Main Result

Main Theorem: Flatness Condition on Features

Natural RL algorithms converge if and only if all linear combinations of the features achieve their extreme values on regions of the state space that have non-zero measure.

Linear Function Approximation

- Let $\Phi = \{\phi_1, \ldots, \phi_k\}$ be the chosen feature functions and $\phi(s) = (\phi_1(s), \ldots, \phi_k(s))^\top$.
- Produce a parametrized estimate $\hat{V}$ of V using linear function approximation: $\hat{V}(s) = \sum_{i=0}^{k} \phi_i(s) w_i = \phi(s)^\top w$.

Definition: Flat Extrema

Let φ be any linear combination of the features Φ. Then we say that φ has flat extrema if it achieves its max (and min) values on a region of the state space with non-zero measure.

Example of flat/non-flat maxima

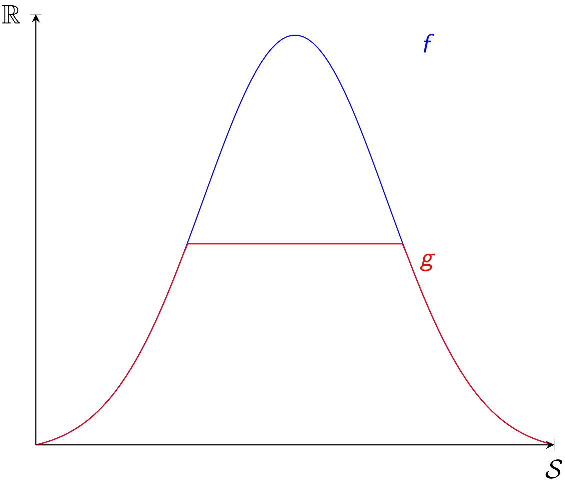


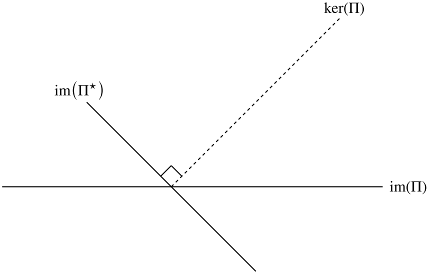
Figure: Function f has a non-flat maxima since it achieves its max value at a point whereas g clearly has a flat maxima.

Oblique Projection Operators

RL algorithms using linear function approximation can be viewed as taking an oblique projection on the Bellman equation.

Definition: Oblique Projection Operators

Let $\Phi = \{\phi_1, ..., \phi_k\}$ and $\Psi = \{\psi_1, ..., \psi_n\}$. Let Π be an oblique projection operator such that $\text{im}(\Pi) = \text{span}(\Phi)$ and $\text{im}(\Pi^*) = \text{span}(\Psi)$. Then Π can be characterised by the two sets (Φ, Ψ).



Geometry of Linear Value Function Approximation

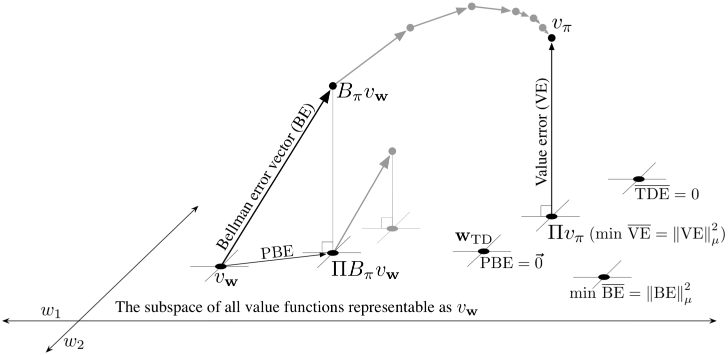


Figure: The geometry of linear value function approximation (Sutton and Barto, 2018). The Bellman operator is given by B_π, v_π is the true value function, v_w is the linear value function approximation, *BE* is the Bellman error, *PBE* is the projected Bellman error, *VE* is the value error, and *TDE* is the TD error.

Geometry of Linear Value Function Approximation continued...

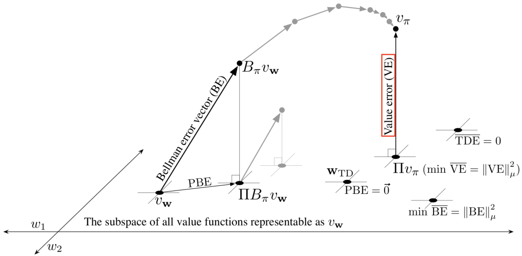


Figure: The geometry of linear value function approximation (Sutton and Barto, 2018).

- Value Error (VE):
  - Best approximation to the true value function v_π.
  - Requires knowledge of v_π, which we don't have.

Geometry of Linear Value Function Approximation continued...

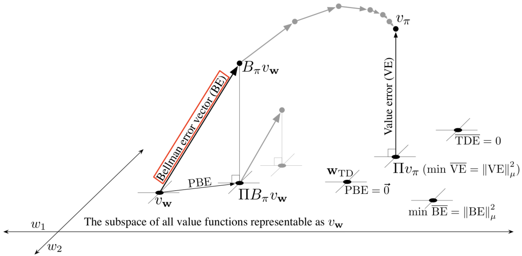


Figure: The geometry of linear value function approximation (Sutton and Barto, 2018).

- Bellman Error (BE):
  - Difference between two sides of Bellman equation.
  - A measure of how far v_w is from v_π.
  - Suffers from non-uniqueness $\implies$ not learnable.

Geometry of Linear Value Function Approximation continued...

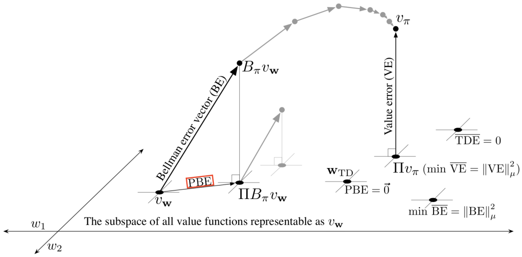


Figure: The geometry of linear value function approximation (Sutton and Barto, 2018).

- Projected Bellman Error (*PBE*):
  - Solution is the *TD*(0) fixed point.
  - Approximation to *BE*.
  - Counter-examples exist showing that this point is not stable under traditional *TD* learning.

Convergence

We analyse non-uniqueness and convergence in the case where $BE = 0$, i.e, the true value function v_π is exactly representable in the approximation subspace.

Definition: Convergence

Let (R^*, T^*) be the true environment with optimal value function V^* that can be represented as $V^* = \phi^\top w^*$. An algorithm is said to converge if it converges to w^*, or equivalently, V^*.

Definition: Failure to Converge

An algorithm is said to fail if there exists an environment $(R^\dagger, T^\dagger)$ with optimal parameter vector $w^\dagger$ such that $w^\dagger \neq w^*$ that it cannot distinguish from (R^*, T^*).

Main Theorem: Flatness Condition on Features

Main Theorem: Flat Extrema Condition

Natural RL algorithms converge if and only if all linear combinations of the features have flat extrema.

Main Theorem: Flatness Condition on Features continued...

- Condition is model-agnostic and policy-agnostic.
- All linear combinations of the features must have the flat extrema property.

Example: State Aggregation

Corollary: State Aggregation

State aggregation is a feature construction choice that will always converge.

Example: State Aggregation

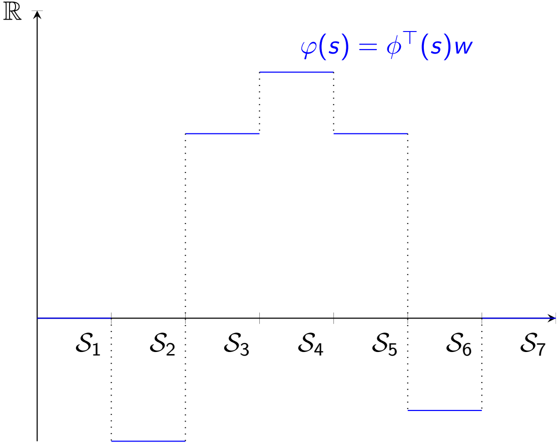


Figure: A partitioning state-aggregation feature construction that generates non-measure zero partitions and hence has flat extrema.

A Projection Perspective

The inner product between ψ_i and any function $\varphi \in \text{span}(\Phi)$ can be seen as the projection of φ onto ψ_i.

Theorem: Projection Condition for Convergence

All algorithms characterized by (Φ, Ψ) converge if and only if for all $\varphi \in \text{span}(\Phi)$ there exists an i such that

$$\langle \psi_i, \varphi \rangle_\mu \geq G\varphi_{\max} \quad \text{or} \quad \langle \psi_i, \varphi \rangle_\mu \leq G\varphi_{\min} \ , \tag{1}$$

where $G := \frac{(1-\lambda)\gamma}{1-\lambda\gamma}$.

A Projection Perspective continued...

- Project on sub-regions of state space that achieve extreme values $\implies$ convergence.
- Simple case: $\langle \psi_i, \varphi \rangle_\mu = \varphi_{\text{max}}$ or $\langle \psi_i, \varphi \rangle_\mu = \varphi_{\text{min}}$.
- Occurs when φ has flat extrema and ψ_i projects on the flat extrema.

Example: Projecting on Flat Extrema

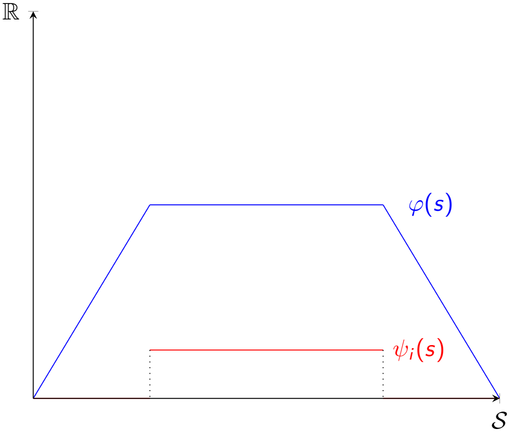


Figure: An example of a convergent algorithm. This algorithm precisely projects on the flat extrema, thus satisfying (1).

Example: Constructing a Convergent Algorithm

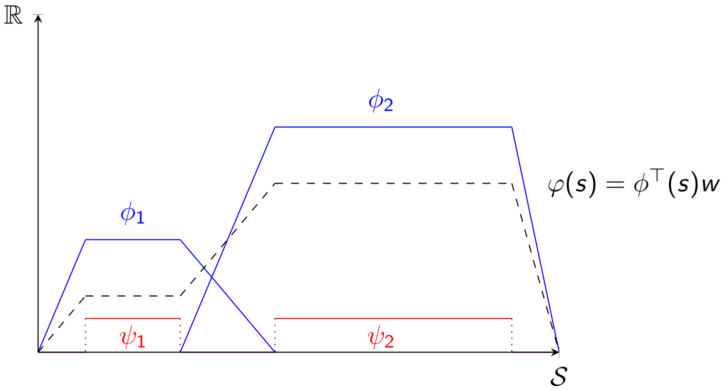


Figure: An example of a convergent algorithm constructed from piece-wise linear features (ϕ_1, ϕ_2) with projection components (ψ_1, ψ_2). Algorithms of this form are guaranteed to converge.

Discussion

- The choice of discount factor plays an important role in defining flat extrema.
- As it moves away from one and towards zero, it becomes less likely that non-flat extrema will occur.
- $\implies$ The discount factor determines the degree to which feature choices that deviate from flat extrema can converge.

Discussion

- Can our work be generalised to consider approximate convergence?
- The natural algorithms do not cover all RL algorithms with linear function approximation. In fact, the ETD algorithm may not fit our framework.

Appendix: Flat Extrema Condition Proof Idea

Lemma

Let $0 \not\equiv \varphi \in \text{span}(\Phi)$ and $\varphi_{\min} := \min_{s \in \mathcal{S}} \varphi(s)$ and $\varphi_{\max} := \max_{s \in \mathcal{S}} \varphi(s)$. Then all natural algorithms fail if and only if there exists an $f : \mathcal{S} \to [\varphi_{\min}, \varphi_{\max}]$ such that

$$\langle \psi_i, \varphi \rangle_\mu = G \langle \psi_i, f \rangle_\mu \tag{2}$$

for all $i = 1, \ldots, n$ and where $G := \frac{(1-\lambda)\gamma}{1-\lambda\gamma}$.

Proof idea:

- Show that such a function f exists if and only if φ has non-flat extrema.
- Taking the contra-positive gives our convergence result.

Appendix: Flat Extrema Condition Proof Idea

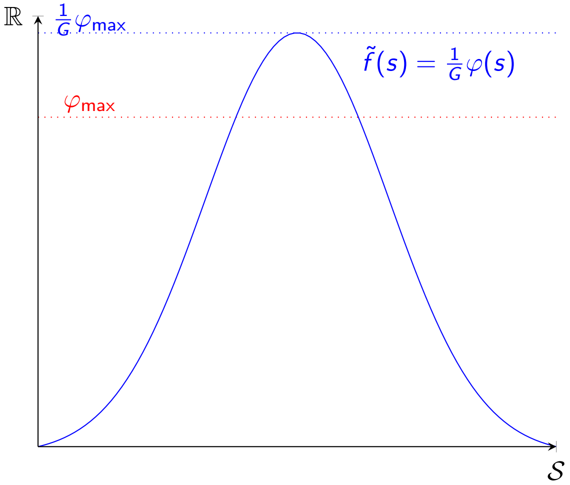


Figure: A function $\tilde{f}$ that exceeds the upper limit on the range but satisfies (1).

Appendix: Flat Extrema Condition Proof Idea

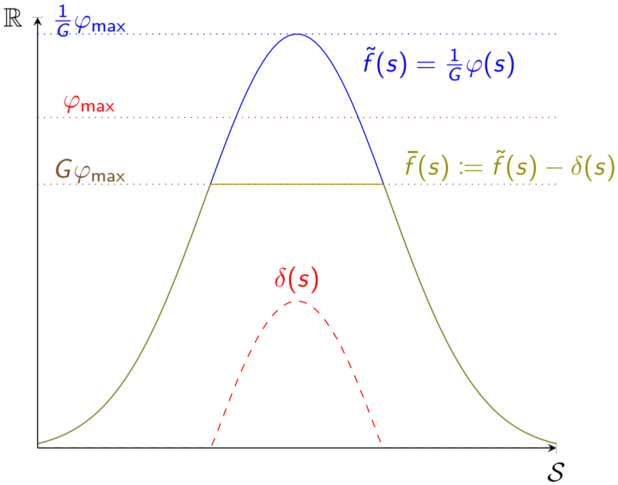


Figure: Define a new function $\bar{f}$ that 'cuts' the top pinnacle that exceeds $G\varphi_{\max}$. The pinnacle δ has 'mass' $o(1-G)$.

Appendix: Flat Extrema Condition Proof Idea

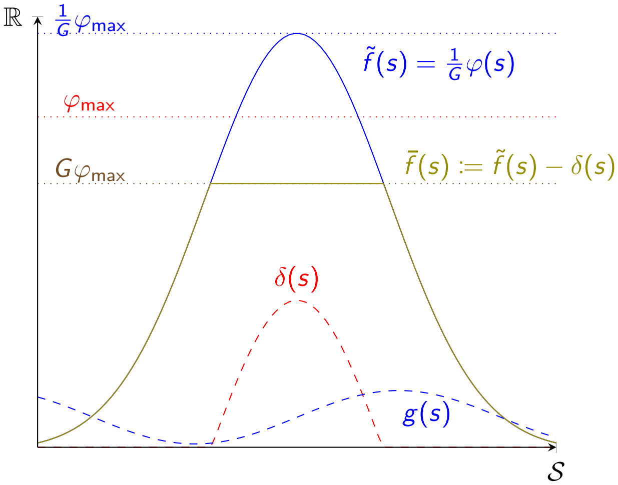


Figure: Define g as the projection of the cut pinnacle δ across the basis functions ψ_i.

Appendix: Flat Extrema Condition Proof Idea

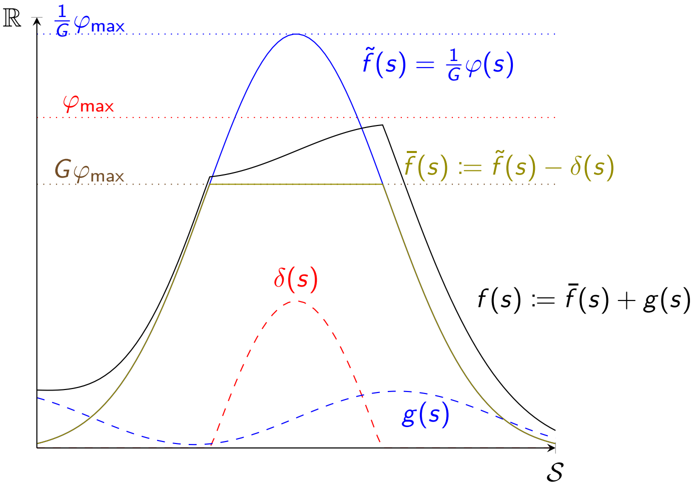


Figure: Define a new function f as the combination of $\bar{f}$ and g. The function f satisfies both the upper bound on the range and (1).

References

Scherrer, Bruno (2010). "Should one compute the Temporal Difference fix point or minimize the Bellman Residual? The unified oblique projection view". In: arXiv:1011.4362.

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. The MIT Press.

Sutton, Richard S. et al. (2009). "Fast Gradient Descent Methods for Temporal-Difference Learning With Linear Function Approximation". In: *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 993–1000.

Tsitsiklis, John N. and Benjamin Van Roy (1997). "An Analysis of Temporal-Difference Learning with Function Approximation". In: *IEEE Transactions on Automatic Control* 42.5, pp. 674–690.