Temporal Difference Learning as Gradient Splitting

Rui Liu [1] Alex Olshevsky [2]

[1]Division of Systems Engineering, Boston University

[2]Department of ECE and Division of Systems Engineering, Boston University

BOSTON UNIVERSITY

Markov Decision Processes (MDP)

- We consider a discounted reward MDP described by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$
  - $\mathcal{S}$: finite state space; $\mathcal{A}$: finite action space; $\mathcal{P}$: transition probabilities; r: rewards; γ: discount factor.

Markov Decision Processes (MDP)

- We consider a discounted reward MDP described by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$
  - $\mathcal{S}$: finite state space; $\mathcal{A}$: finite action space; $\mathcal{P}$: transition probabilities; r: rewards; γ: discount factor.
- Value function of a given stationary policy μ:

$$V^{\mu}(s) = E_{\mu,s}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1}\right]$$

Markov Decision Processes (MDP)

- We consider a discounted reward MDP described by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$
  - $\mathcal{S}$: finite state space; $\mathcal{A}$: finite action space; $\mathcal{P}$: transition probabilities; r: rewards; γ: discount factor.
- Value function of a given stationary policy μ:

$$V^{\mu}(s) = E_{\mu,s}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1}\right]$$

- Policy evaluation refers to the problem of estimating the value function V^{μ}.

Assumption on Markov Chain

- For a given stationary policy μ, the probability transition matrix P^μ can be defined as:

$$P^\mu(s,s') = \sum_{a\in\mathcal{A}} \mu(s,a)\mathcal{P}(s'|s,a).$$

Assumption on Markov Chain

- For a given stationary policy μ, the probability transition matrix P^μ can be defined as:

$$P^\mu(s,s') = \sum_{a \in \mathcal{A}} \mu(s,a)\mathcal{P}(s'|s,a).$$

Assumption 1

The Markov chain whose transition matrix is the matrix P^μ is irreducible and aperiodic.

Assumption on Markov Chain

- For a given stationary policy μ, the probability transition matrix P^μ can be defined as:

$$P^\mu(s, s') = \sum_{a \in \mathcal{A}} \mu(s, a) \mathcal{P}(s'|s, a).$$

Assumption 1

The Markov chain whose transition matrix is the matrix P^μ is irreducible and aperiodic.

- Following this assumption, the Markov decision process induced by the policy μ is ergodic with a unique stationary distribution $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$

Linear Function Approximation

- To reduce computational complexity, a standard remedy is to use low dimensional approximation V_θ^μ of V^μ in the classical TD algorithm.

Linear Function Approximation

- To reduce computational complexity, a standard remedy is to use low dimensional approximation V_θ^μ of V^μ in the classical TD algorithm.
- Consider linear function approximation:

$$V_\theta^\mu(s) = \sum_{l=1}^{K} \theta_l \phi_l(s) \quad \forall s \in \mathcal{S}$$

for a given set of K feature vectors $\phi_l : \mathcal{S} \to \mathbb{R}$, $l \in [K]$. Furthermore, let

$$\phi(s) = (\phi_1(s), \phi_2(s), \cdots, \phi_K(s))^T \in \mathbb{R}^K.$$

Linear Function Approximation

- To reduce computational complexity, a standard remedy is to use low dimensional approximation V_θ^μ of V^μ in the classical TD algorithm.
- Consider linear function approximation:

$$V_\theta^\mu(s) = \sum_{l=1}^{K} \theta_l \phi_l(s) \quad \forall s \in \mathcal{S}$$

  for a given set of K feature vectors $\phi_l : \mathcal{S} \to \mathbb{R}$, $l \in [K]$. Furthermore, let

$$\phi(s) = (\phi_1(s), \phi_2(s), \cdots, \phi_K(s))^T \in \mathbb{R}^K.$$

Assumption 2

The feature vectors $\{\phi_1, \ldots, \phi_K\}$ are linearly independent. Additionally, we also assume that $\|\phi(s)\|_2^2 \leq 1$ for $s \in \mathcal{S}$.

TD(0) with Linear Function Approximation

- TD(0) with linear function approximation updates parameter vector as:

$$\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t),$$

where $g_t(\theta_t) = (r(s_t, s_t') + \gamma \theta_t^T \phi(s_t') - \theta_t^T \phi(s_t))\phi(s_t)$.

TD(0) with Linear Function Approximation

- TD(0) with linear function approximation updates parameter vector as:

$$\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t),$$

where $g_t(\theta_t) = (r(s_t, s_t') + \gamma \theta_t^T \phi(s_t') - \theta_t^T \phi(s_t))\phi(s_t)$.

- Let $\bar{g}(\theta)$ denote the average of $g_t(\theta)$:

$$\bar{g}(\theta) = \sum_{s,s' \in \mathcal{S}} \pi(s) P^{\mu}(s,s') \left(r(s,s') + \gamma \phi(s')^T \theta - \phi(s)^T \theta \right) \phi(s).$$

TD(0) with Linear Function Approximation

- TD(0) with linear function approximation updates parameter vector as:

$$\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t),$$

  where $g_t(\theta_t) = (r(s_t, s_t') + \gamma \theta_t^T \phi(s_t') - \theta_t^T \phi(s_t))\phi(s_t)$.
- Let $\bar{g}(\theta)$ denote the average of $g_t(\theta)$:

$$\bar{g}(\theta) = \sum_{s,s' \in \mathcal{S}} \pi(s) P^\mu(s,s') \left(r(s,s') + \gamma \phi(s')^T \theta - \phi(s)^T \theta \right) \phi(s).$$

- Under Assumptions 1-2 as well as an additional assumption on the decay of the step-sizes α_t, TD learning converges almost surely; furthermore, its limit θ^* satisfies: $\bar{g}(\theta^*) = 0$. [Tsitsiklis & Van Roy(1997)]

Gradient Splitting and Gradient Descent

Definition of Gradient Splitting

Let A be a symmetric positive semi-definite matrix. A linear function $h(\theta) = B(\theta - a)$ is called a gradient splitting of the quadratic $f(\theta) = (\theta - a)^T A(\theta - a)$ if

$$B + B^T = 2A.$$

Gradient Splitting and Gradient Descent

Definition of Gradient Splitting

Let A be a symmetric positive semi-definite matrix. A linear function $h(\theta) = B(\theta - a)$ is called a gradient splitting of the quadratic $f(\theta) = (\theta - a)^T A(\theta - a)$ if

$$B + B^T = 2A.$$

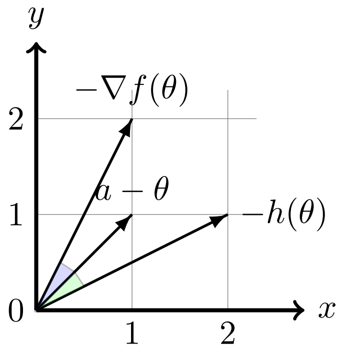
Proposition 1 [Why is gradient splittings useful?]

Suppose $h(\theta)$ is a splitting of the gradient of $f(\theta)$. Then

$$(\theta_1 - \theta_2)^T (h(\theta_1) - h(\theta_2)) = \frac{1}{2}(\theta_1 - \theta_2)^T (\nabla f(\theta_1) - \nabla f(\theta_2)).$$

Furthermore, for all θ, $(a - \theta)^T h(\theta) = \frac{1}{2}(a - \theta)^T \nabla f(\theta)$.

Example



- $\theta = (0,0)^T$, $a = (1,1)^T$, $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix}$
- $f(\theta) = (\theta - a)^T A(\theta - a)$, $h(\theta) = B(\theta - a)$. $h(\theta)$ is a gradient splitting of $f(\theta)$.

More Comments

- Negative gradient splitting has the same positive inner product with the direction to optimality as the negative gradient.
- Therefore, gradient splitting "makes progress" towards the optimal solution as gradient descent.
- As a consequence of this discussion, we can apply the existing proof for gradient descent almost verbatim to gradient splittings.

Mean-path TD(0)

- Mean-path TD(0) updates parameter vector as:

$$\theta_{t+1} = \theta_t + \alpha_t \bar{g}(\theta_t).$$

Mean-path TD(0)

- Mean-path TD(0) updates parameter vector as:

$$\theta_{t+1} = \theta_t + \alpha_t \bar{g}(\theta_t).$$

- Will the mean-path TD update brings θ_t closer to θ^*?
  - $\bar{g}(\theta)^T(\theta^* - \theta) > 0$. [Tsitsiklis & Van Roy(1997)]
  - $\bar{g}(\theta)^T(\theta^* - \theta) \geq (1-\gamma)\|V_{\theta^*} - V_\theta\|_D^2$ [Tsitsiklis & Van Roy(1997), Bhandari et al(2018)], where

$$\|V\|_D^2 = V^T D V = \sum_{s \in \mathcal{S}} \pi_s V^2(s).$$

Our Main Result

Theorem 1

Suppose Assumptions 1-2 hold. Then in the TD(0) update, $-\bar{g}(\theta)$ is a splitting of the gradient of the quadratic

$$f(\theta) = (1-\gamma)\| V_\theta - V_{\theta^*} \|_D^2 + \gamma \| V_\theta - V_{\theta^*} \|_{\mathrm{Dir}}^2,$$

where $\| V \|_{\mathrm{Dir}}^2 = \frac{1}{2} \sum_{s,s' \in \mathcal{S}} \pi_s P(s,s')(V(s') - V(s))^2$.

Our Main Result

Theorem 1

Suppose Assumptions 1-2 hold. Then in the TD(0) update, $-\bar{g}(\theta)$ is a splitting of the gradient of the quadratic

$$f(\theta) = (1-\gamma)\| V_\theta - V_{\theta^*} \|_D^2 + \gamma \| V_\theta - V_{\theta^*} \|_{\mathrm{Dir}}^2,$$

where $\| V \|_{\mathrm{Dir}}^2 = \frac{1}{2} \sum_{s,s' \in \mathcal{S}} \pi_s P(s,s')(V(s') - V(s))^2$.

Corollary 1

For any $\theta \in \mathbb{R}^K$,

$$(\theta^* - \theta)^T \bar{g}(\theta) = (1-\gamma)\| V_{\theta^*} - V_\theta \|_D^2 + \gamma \| V_{\theta^*} - V_\theta \|_{\mathrm{Dir}}^2.$$

Markovian Samples and Step-size

- We want use Corollary 1 to obtain improved convergence times for TD(0).

Markovian Samples and Step-size

- We want use Corollary 1 to obtain improved convergence times for TD(0).
- Collecting data: a single sample path of a Markov chain.

Markovian Samples and Step-size

- We want use Corollary 1 to obtain improved convergence times for TD(0).
- Collecting data: a single sample path of a Markov chain.
- Choice of step-size: $O(1/\sqrt{T})$
  - For faster decaying step-sizes, for example $O(1/t)$, performance will scale with the inverse of the smallest eigenvalue of $\Phi^T D \Phi$ or related quantity, and these can be quite small.
  - However, for step-size $O(1/\sqrt{T})$, this is not the case.

Assumption on Markovian Samples

Assumption 3

There are constants $m > 0$ and $\rho \in (0, 1)$ such that

$$\sup_{s \in \mathcal{S}} d_{\mathrm{TV}}(P^t(s, \cdot), \pi) \le m\rho^t \quad t \in \mathbb{N}_0,$$

where $d_{\mathrm{TV}}(P, Q)$ denotes the total-variation distance between probability measures P and Q. In addition, the initial distribution of s_0 is the steady-state distribution π, so that $(s_0, s_1, \cdots)$ is a stationary sequence.

Projected TD(0)

- Consider the projected TD(0) update:

$$\theta_{t+1} = \mathrm{Proj}_{\Theta}(\theta_t + \alpha_t g_t(\theta_t)),$$

where Θ is a convex set containing the optimal solution θ^*.

Projected TD(0)

- Consider the projected TD(0) update:

$$\theta_{t+1} = \mathrm{Proj}_{\Theta}(\theta_t + \alpha_t g_t(\theta_t)),$$

  where Θ is a convex set containing the optimal solution θ^*.
- Moreover, we will assume that the norm of every element in Θ is at most R_θ.

Improved Error Bounds

Corollary 2

Suppose Assumptions 1-3 hold. Suppose further that $(\theta_t)_{t\geq 0}$ is generated by the Projected TD algorithm with $\theta^* \in \Theta$ and $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$. Then

$$E\left[(1-\gamma)\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 + \gamma\|V_{\theta^*} - V_{\bar{\theta}_T}\|_{\mathrm{Dir}}^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left[9 + 12\tau^{\mathrm{mix}}\left(1/\sqrt{T}\right)\right]}{2\sqrt{T}},$$

where τ^{mix} is standard notation for the mixing time of the Markov chain: $\tau^{\mathrm{mix}}(\varepsilon) = \min\left\{t \in \mathbb{N}, t \geq 1 | m\rho^t \leq \varepsilon\right\}$.

Improved Error Bounds

Corollary 2

Suppose Assumptions 1-3 hold. Suppose further that $(\theta_t)_{t\geq 0}$ is generated by the Projected TD algorithm with $\theta^* \in \Theta$ and $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$. Then

$$E\left[(1-\gamma)\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 + \gamma\|V_{\theta^*} - V_{\bar{\theta}_T}\|_{\mathrm{Dir}}^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left[9 + 12\tau^{\mathrm{mix}}\left(1/\sqrt{T}\right)\right]}{2\sqrt{T}},$$

where τ^{mix} is standard notation for the mixing time of the Markov chain: $\tau^{\mathrm{mix}}(\varepsilon) = \min\left\{t \in \mathbb{N}, t \geq 1 | m\rho^t \leq \varepsilon\right\}$.

- We also generalize gradient splitting and improved error bound on TD(0) to TD(λ) in our paper.

Compare to Existing Bounds

- Theorem 3(a) in Bhandari et al(2018):

$$E\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{2(1-\gamma)\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\mathrm{mix}}\left(1/\sqrt{T}\right)\right]}{2(1-\gamma)\sqrt{T}}.$$

Compare to Existing Bounds

- Theorem 3(a) in Bhandari et al(2018):

$$E\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{2(1-\gamma)\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2(1-\gamma)\sqrt{T}}.$$

- This upper bound blows up as $\gamma \to 1$.

Compare to Existing Bounds

- Theorem 3(a) in Bhandari et al(2018):

$$E\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{2(1-\gamma)\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2(1-\gamma)\sqrt{T}}.$$

- This upper bound blows up as $\gamma \to 1$.
- However, based on Corollary 2, we can obtain

$$E\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_{\text{Dir}}^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{2\gamma\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2\gamma\sqrt{T}}.$$

Compare to Existing Bounds

- Theorem 3(a) in Bhandari et al(2018):

$$E\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{2(1-\gamma)\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2(1-\gamma)\sqrt{T}}.$$

- This upper bound blows up as $\gamma \to 1$.
- However, based on Corollary 2, we can obtain

$$E\left[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_{\text{Dir}}^2\right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{2\gamma\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2\gamma\sqrt{T}}.$$

- Therefore, the error of averaged & projected temporal difference learning projected on $\mathbf{1}^\perp$ does not blow up as $\gamma \to 1$.

The Scaling with the Discount Factor

- Is it possible to remove the dependence on $O(1/(1-\gamma))$ from bounds on the performance of temporal difference learning?

The Scaling with the Discount Factor

- Is it possible to remove the dependence on $O(1/(1-\gamma))$ from bounds on the performance of temporal difference learning?
- Unfortunately, the answer is no. However, it is possible to derive a bound where the only scaling with $1/(1-\gamma)$ is in the asymptotically negligible term.

Mean-adjusted TD(0)

Algorithm 1 Mean-adjusted TD(0)

Initialize $\bar{A}_0 = 0$, $s_0 \sim \pi$, and some initial condition θ_0.
for $t = 0$ to $T-1$ **do**
    Projected TD(0) update:
    $\theta_{t+1} = \mathrm{Proj}_{\Theta}(\theta_t + \alpha_t g_t(\theta_t))$
    Keep track of the average reward: $\bar{A}_{t+1} = \frac{t\bar{A}_t + r_{t+1}}{t+1}$
end for
Set $\hat{V}_T = \frac{\bar{A}_T}{1-\gamma}$
Output $V'_T = V_{\bar{\theta}_T} + \left(\hat{V}_T - \pi^T V_{\bar{\theta}_T}\right) 1$

A Better Scaling with the Discount Factor

Corollary 3

Suppose that $(\theta_t)_{t\geq 0}$ and V'_T are generated by Algorithm 1 with step-sizes $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$. Let t_0 be the largest integer which satisfies $t_0 \leq 2\tau^{\mathrm{mix}}\left(\frac{1}{2(t_0+1)}\right)$. Then as long as $T \geq t_0$, we will have

$$E\left[\|V'_T - V\|_D^2\right] \leq O\left(E\left[\|V_{\theta^*} - V\|_D^2\right] + \frac{r_{\max}^2 \tau^{\mathrm{mix}}\left(\frac{1}{2(T+1)}\right)}{(1-\gamma)^2 T}\right.$$
$$\left. + \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left[1 + \tau^{\mathrm{mix}}(1/\sqrt{T})\right]}{\sqrt{T}} \min\left\{\frac{r(P)}{\gamma}, \frac{1}{1-\gamma}\right\}\right).$$