TD(0) with linear function approximation

Csaba Szepesvári

Wednesday 29th March, 2023

On-policy TD(0), linearly approximated,
In reinforcement learning, it's highly rated.
Updating weights with each new step,
Learning with bootstrapping, never to forget.

With every move, it seeks to improve,
Converging to some good groove.
Value estimation is the key,
To finding the best policy.

The algorithm's simplicity,
Makes it a favorite of the RL community.
A poem to honor its efficacy,
On-policy TD(0) with linear function approximation, oh so mighty.
– ChatGPT with some edits from yours' truly

1 The problem

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ be a finite MDP, with $\mathcal{S} = \{1, \dots, S\}$. Let $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{S})$ be a memoryless stochastic policy. Let $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \dots$ be a trajectory. We let $\mathbb{P}$ denote the probability distribution over the sample space that holds $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \dots$ that arises from following π in $\mathcal{M}$ starting from a distribution $\mu \in \mathcal{M}_1(\mathcal{S})$, that is, $S_0 \sim \mu$ and for $t \ge 0$, $A_t \sim \pi(\cdot|S_t)$ and $(R_{t+1}, S_{t+1}) \sim p(\cdot, \cdot|S_t, A_t)$. Further, we assume we are given a map $x : \mathcal{S} \to \mathbb{R}^d$, which we will call a **feature-map**. For a vector $\theta \in \mathbb{R}^d$, let $v_\theta : \mathcal{S} \to \mathbb{R}$ be defined by

$$v_\theta(s) = \theta^\top x(s)\,, \qquad s \in \mathcal{S}.$$

Following the standard nomenclature in the literature, we call θ a **weight vector** (its components are used in a weighted sum together with the components of the features to get a value). The **problem considered** is to design an incremental update rule that generates a sequence of vectors $(\theta_t)_{t\ge0}$ such that $\theta_t \to \theta_\infty$ in an appropriate sense and such that v_{θ_∞} is close to v_π.

2 TD(0) with linear function approximation

Given $\theta_0 \in \mathbb{R}^d$, the feature map x, the above data, and a steps-size sequence $(\alpha_t)_{t\geq 0}$ of non-negative numbers, TD(0) generates a sequence of vectors $(\theta_t)_{t\geq 1}$ such that for all $t \geq 0$,

$$\theta_{t+1} = \theta_t + \alpha_t \delta_t(\theta_t) X_t\,, \tag{1}$$

where we define X_t to be the feature vector of the state S_t visited at time t:

$$X_t = x(S_t)\,.$$

Further, we define $\delta_t(\theta)$, the temporal difference error, or, in short, the TD-error, at time t when the weight vector θ is used, as

$$\delta_t(\theta) = R_{t+1} + \gamma v_\theta(S_{t+1}) - v_\theta(S_t)\,.$$

3 Convergence

The sequence of states, $(S_t)_{t\geq 0}$ forms a Markov chain[1] and for any t, $s, s' \in \mathcal{S}$ such that $\mathbb{P}(S_t = s) > 0$, $\mathbb{P}(S_{t+1} = s'|S_t = s) = (P_\pi)_{s,s'}$ where the transition matrix $P_\pi \in [0,1]^{S\times S}$ is given by

$$(P_\pi)_{s,s'} = \sum_{a\in\mathcal{A}} \pi(a|s)p(s'|s,a)\,.$$

A distribution $\mu \in \mathcal{M}_1(\mathcal{S})$ is called a **stationary distribution** of P_π, if for all $s' \in \mathcal{S}$ it holds that

$$\sum_{s\in\mathcal{S}} \mu(s)(P_\pi)_{s,s'} = \mu(s')\,.$$

It follows that if the distribution of S_t is μ (that is, $\mathbb{P}(S_t = s) = \mu(s)$, $s \in \mathcal{S}$) then S_{t+1} also has the distribution μ. Every finite Markov chain has at least one stationary distribution.[2] Thinking of μ as a row-vector (which is how we, conventionally, think of distributions when we view them as vectors), the condition that μ is a stationary distribution of P_π takes the form

$$\mu P_\pi = \mu.$$

We call a Markov chain **irreducible** if every state can be reached from every other state. If a Markov chain is not irreducible, its state space can be partitioned into multiple non-empty parts such that states within the same part can be reached from each other, but states belonging to different parts cannot be reached from each other. A finite state space Markov chain that is not irreducible has more than one stationary distribution.

We call a state $s \in \mathcal{S}$ **periodic** with period $t \geq 2$ if for any $n > 0$, $(P_\pi^n)_{s,s}$ is non-zero if and only if n is an integer multiple of t. Note that here $(P_\pi^n)_{s,s'}$ gives the probability that the chain will get to state s' in n steps when it starts at state s. A Markov chain that has no periodic states is called **aperiodic**. A simple example of a Markov chain that is not aperiodic is a chain with two states, where the transitions are such that the chain alternates (deterministically) between the two states. The period of both states in this case is two.

A Markov chain that is both irreducible and aperiodic has nice properties. In particular, it has a unique stationary distribution and no matter what the initial distribution of the states is, the distribution $\mu_t(s) = \mathbb{P}(S_t = s)$ for all $s \in \mathcal{S}$, converges to the chain's unique stationary distribution. Further, this convergence is exponentially fast: $\|\mu_t - \mu\| \le C\exp(-ct)$ for all $t \ge 0$, where $c, C > 0$ are fixed positive constants.

The following theorem holds true:

Theorem 3.1 (Convergence of TD(0))**.** *Assume that $(S_t)_{t\ge 0}$ defines an irreducible and aperiodic Markov chain with stationary distribution $\mu \in \mathcal{M}_1(\mathcal{S})$. Assume further that $\alpha_t \ge 0$, $\sum_t \alpha_t = \infty$, and $\sum_t \alpha_t^2 < \infty$ and that the matrix $G = \sum_{s\in\mathcal{S}} \mu(s)x(s)x(s)^\top$ is non-singular. Then there exists $\theta_* \in \mathbb{R}^d$ such that $\theta_t \to \theta_*$ with probability one.*

We note in passing that the condition that the Markov chain is aperiodic is for convenience. As we shall see this helps with some of the arguments. However, the claim would continue to hold even without this assumption. We will not prove this theorem, but will sketch the main reasons behind its convergence. However, we first discuss the properties of θ_*.

4 On the limit point(s) of TD(0)

Let $\mu_t(s) = \mathbb{P}(S_t = s)$, for $t \ge 0$. Define

$$f_t(\theta) = \mathbb{E}[\delta_t(\theta)X_t]\,.$$

A simple calculation then shows that

$$f_t(\theta) = \sum_{s\in\mathcal{S}} \mu_t(s)x(s)((T_\pi v_\theta)(s) - v_\theta(s))\,.$$

As noted earlier, if (P_π) is irreducible and aperiodic, $\mu_t \to \mu$, where μ is the unique stationary distribution of P_π. Hence, assuming that $\theta_t \to \theta_*$ as $t \to \infty$, we expect that

$$f(\theta_*) = 0 \tag{2}$$

must hold, where

$$f(\theta) = \sum_{s\in\mathcal{S}} \mu(s)x(s)((T_\pi v_\theta)(s) - v_\theta(s))\,. \tag{3}$$

In the rest of this section we examine the consequences of Equation (2). For this, we will only need that μ is a stationary distribution of P_π, but we will not need to assume that P_π is irreducible and aperiodic.

Define a family of functions $(J_\theta)_{\theta\in\mathbb{R}^d}$ as follows:

$$J_\theta(\theta') = \frac{1}{2}\|T_\pi v_\theta - v_{\theta'}\|_\mu^2\,.$$

Here, for a function $u : \mathcal{S} \to \mathbb{R}$ we let

$$\|u\|_\mu^2 = \sum_{s\in\mathcal{S}} \mu(s)u^2(s)$$

to be the μ-weighted squared two-norm of u. A simple calculation shows that

$$f(\theta) = -\nabla J_\theta(\theta)\,. \tag{4}$$

Indeed, fixing θ and letting $u = T_\pi v_\theta$, we have

$$\begin{aligned}
\nabla J_\theta(\theta') &= \frac{1}{2}\nabla_{\theta'} \sum_s \mu(s)(v_{\theta'}(s) - u(s))^2 \\
&= \sum_s \mu(s)(v_{\theta'}(s) - u(s))\nabla_{\theta'} v_{\theta'}(s) \\
&= \sum_s \mu(s)(v_{\theta'}(s) - u(s))x(s)\,. \qquad (5)
\end{aligned}$$

Plugging back $u = T_\pi v_\theta$ and substituting θ for θ' gives the result.[3]

Now, from Equation (4) it follows that $f(\theta_*) = 0$ implies that $\nabla J_{\theta_*}(\theta_*) = 0$. On the other hand, for $\theta, \theta' \in \mathbb{R}^d$, $\nabla J_\theta(\theta') = 0$ implies that

$$\theta' \in \arg\min_{\theta''} J_\theta(\theta'')\,.$$

This holds because J_θ is a convex function of its argument and convex functions have the property that if a point makes the function's derivative (or gradient) zero then it must be a minimizer of the function. It follows that

$$\theta_* \in \arg\min_{\theta''} J_{\theta_*}(\theta'')\,. \tag{6}$$

Define the **feature matrix** $X \in \mathbb{R}^{S\times d}$, so that the sth row of X is $x^\top(s)$. Note that if we identify v_θ with a vector, as usual, the equality $v_\theta = X\theta$ holds for any $\theta \in \mathbb{R}^d$.

In what follows, **we assume** that the $d \times d$ matrix $G = \sum_s \mu(s)x(s)x(s)^\top$ that was defined in Theorem 3.1, and which can be equivalently defined via

$$G = X^\top D X,$$

is invertible. Here, $D = \mathrm{diag}(\mu)$ is the diagonal matrix whose diagonal entries are given by the respective elements of μ.

That G is non-singular implies that for any $u \in \mathbb{R}^S$ the minimizer of

$$\theta \mapsto \|u - v_\theta\|_\mu^2$$

is uniquely defined (you might recall this argument, from a previous ML course, used while discussing weighted least squares regression). This in turns allows us to define the **projection** operator $\Pi : \mathbb{R}^S \to \mathbb{R}^S$, which is defined as follows:

$$\Pi u = X \arg\min_{\theta\in\mathbb{R}^d} \|u - v_\theta\|_\mu^2\,. \tag{7}$$

An equivalent definition of Π is that it gives the unique $v = \Pi u \in \mathbb{R}^d$ such that

$$\|u - v\|_\mu^2 = \min_{\theta\in\mathbb{R}^d} \|u - v_\theta\|_\mu^2\,.$$

By some calculation, we can also see that

$$\Pi u = XG^{-1}X^\top Du\,,$$

and from this we see that Π is a linear operator. We can thus also identify Π with the matrix $XG^{-1}X^\top D$. From this, we also immediately see that for any $\theta \in \mathbb{R}^d$, $u \in \mathbb{R}^S$,

$$\langle v_\theta, u - \Pi u\rangle_\mu = 0\,, \tag{8}$$

where we use the notation

$$\langle u, v\rangle_\mu = \sum_{s\in\mathcal{S}} \mu(s)u(s)v(s) = u^\top Dv\,.$$

Note that Equation (8) is known as a **Pythagorean identity**: It states that v_θ is orthogonal to $u - \Pi u$ in the geometry induced by $\langle\cdot,\cdot\rangle_\mu$, where we call u and v orthogonal to each other in this geometry exactly when $\langle u, v\rangle_\mu = 0$. In words, this identity expresses the fact that the "error" $u - \Pi u$ induced by approximating u with Πu is orthogonal to any element of the subspace

$$H = \{v_\theta \,:\, \theta \in \mathbb{R}^d\} \tag{9}$$

of $\mathbb{R}^S$.

Now, one can rewrite Equation (6) in terms of v_{θ_*}. Noting that $\theta_* \in \arg\min_{\theta''} J_{\theta_*}(\theta'')$, it follows from the definition of J_{θ_*} (that is, $J_{\theta_*}(\theta') = \frac{1}{2}\|T_\pi v_{\theta_*} - v_{\theta'}\|_\mu^2$) that

$$v_{\theta_*} = \Pi T_\pi v_{\theta_*}\,.$$

Hence, if TD(0) converges, it must converge to a fixed point of the composite operator ΠT_π.

The first question then is whether ΠT_π has a fixed point at all (and whether it has multiple fixed points). As it turns out, the following holds:

Theorem 4.1. *Assume that G is non-singular (so that Π is well-defined). Then,* (i) *ΠT_π is a $\|\cdot\|_\mu$ contraction with contraction factor γ, and* (ii) *it has a unique fixed point.*

Proof. The proof of Part *(i)* consists of showing that the projection Π is a non-expansion, while T_π is a γ-contraction with respect to the chosen norm. Assume now that we have already proven Part *(i)*. Now, consider the vector space H defined by Equation (9). Let $F = \Pi T_\pi$. Clearly, for any $v \in H$, $Fv \in H$. Further, $\|\cdot\|_\mu$ on H is a norm: Clearly, $\|\cdot\|_\mu$ satisfies the triangle inequality, it is non-negative and positive homogeneous, even when restricted to H. Further, for $v \in H$, it also holds that $\|v\|_\mu = 0$ implies that $v = 0$. Indeed, from $v \in H$ it follows that $v = v_\theta$ for some $\theta \in \mathbb{R}^d$. Hence, $\|v\|_\mu^2 = \|v_\theta\|_\mu^2 = \theta^\top G\theta$. This, if $\|v\|_\mu^2 = 0$ then $\theta^\top G\theta = 0$. Since G is non-singular, it is positive definitive, hence, $\theta^\top G\theta = 0$ implies that $\theta = 0$, which in turn implies that $v = X0 = 0$. Thus, we can apply Banach's fixed point theorem with the operator F restricted to H and using the norm $\|\cdot\|_\mu$ on H to get that F has a unique fixed point in H. But any fixed point of F needs to be an element of H (since Π maps any vector to H). Hence, we conclude that F has a unique fixed point.

Thus, it remains to prove Part *(i)*. For this, first let us prove that Π is a non-expansion. Abbreviate $\|\cdot\|_\mu$ ($\langle\cdot,\cdot\rangle_\mu$) to $\|\cdot\|$ (respectively, to $\langle\cdot,\cdot\rangle$). Note that $\|u\|^2 = \langle u, u\rangle$. Note also that $\langle u, v\rangle = \langle v, u\rangle$ and $u \mapsto \langle u, v\rangle$ is linear.

Take $u \in \mathbb{R}^S$. Then,

$$\begin{aligned}\|u\|^2 &= \|u - \Pi u + \Pi u\|^2 = \langle v + \Pi u, v + \Pi u \rangle = \langle v, v \rangle + \langle \Pi u, \Pi u \rangle + 2\langle v, \Pi u \rangle \\ &= \|v\|^2 + \|\Pi u\|^2 + 2\langle v, \Pi u \rangle\,,\end{aligned}$$

where we introduced $v = u - \Pi u$. The definition of Π implies that $\Pi u = v_\theta$ for some $\theta \in \mathbb{R}^d$, hence, by Equation (8), $\langle v, \Pi u \rangle = 0$. Therefore,

$$\|\Pi u\|^2 = \|u\|^2 - \|v\|^2 \leq \|u\|^2\,, \tag{10}$$

since $\|v\|^2 \geq 0$. Now, for $u, v \in \mathbb{R}^S$, using that Π is a linear operator and Equation (10), we get

$$\|\Pi u - \Pi v\| = \|\Pi(u - v)\| \leq \|u - v\|\,,$$

which shows that Π is indeed a non-expansion.

Now, for T_π, we have for any $u, v \in \mathbb{R}^S$ that

$$T_\pi v - T_\pi u = r_\pi + \gamma P_\pi v - (r_\pi + \gamma P_\pi u) = \gamma P_\pi (v - u)\,.$$

Hence, if we show that the map $u \mapsto P_\pi u$ is a non-expansion, we will be done. Take any $u \in \mathbb{R}^S$. We have

$$\begin{aligned}\|P_\pi u\|^2 &= \sum_{s \in \mathcal{S}} \mu(s) \Big(\sum_{s' \in \mathcal{S}} (P_\pi)_{s,s'} u(s') \Big)^2 \\ &\leq \sum_{s \in \mathcal{S}} \mu(s) \sum_{s' \in \mathcal{S}} (P_\pi)_{s,s'} u^2(s') \\ &= \sum_{s' \in \mathcal{S}} \Big(\sum_{s \in \mathcal{S}} \mu(s) (P_\pi)_{s,s'} \Big) u^2(s') \\ &= \sum_{s' \in \mathcal{S}} \mu(s') u^2(s') \qquad \text{(because } \mu \text{ is a stationary distribution of } P_\pi\text{)} \\ &= \|u\|^2\,.\end{aligned}$$

Above, the inequality follows from Jensen's inequality, which states that for any convex function $f : \mathbb{R} \to \mathbb{R}$, any function $u : \mathcal{S} \to \mathbb{R}$, and any probability vector $(p(s))_{s \in \mathcal{S}}$, $f(\sum_{s \in \mathcal{S}} p(s) u(s)) \leq \sum_{s \in \mathcal{S}} p(s) f(u(s))$, which we apply here with $f(x) = x^2$ and $p(s') = (P_\pi)_{s,s'}$, $s' \in \mathcal{S}$.

□

From this result, we also get the following result that tells us about the quality of the limit point v_{θ_*}:

Theorem 4.2. *Assume that G is non-singular. Then,*

$$\|v_{\theta_*} - v_\pi\|_\mu \leq \frac{\|\Pi v_\pi - v_\pi\|_\mu}{1 - \gamma}\,.$$

Proof. We have

$$\begin{aligned}v_{\theta_*} - v_\pi &= \Pi T_\pi v_{\theta_*} - T_\pi v_\pi \\ &= (\Pi T_\pi v_{\theta_*} - \Pi T_\pi v_\pi) + (\Pi T_\pi v_\pi - T_\pi v_\pi)\,.\end{aligned}$$

Taking the norm of both sides and using the triangle inequality and that ΠT_π is a γ-contraction, we get

$$\|v_{\theta_*} - v_\pi\|_\mu \leq \gamma \|v_{\theta_*} - v_\pi\|_\mu + \|\Pi v_\pi - v_\pi\|_\mu\,.$$

Solving the inequality gives the result. □

5 The convergence of TD(0)

We can write the update equation of TD(0), given by Equation (1), as

$$\theta_{t+1} = \theta_t + \alpha_t f(\theta_t) + \alpha_t (f_t(\theta_t) - f(\theta_t)) + \alpha_t(\delta_t(\theta_t) X_t - f_t(\theta_t))\,.$$

Here, the error term $f_t(\theta_t) - f(\theta)$ is expected to be "well-behaved" when θ_t does not grow too fast (or, even better, if it remains bounded) since $|f_t(\theta) - f(\theta)| \le C\|\mu - \mu_t\|_1(1 + \|\theta\|_2)$ for some $C > 0$ and since $\|\mu - \mu_t\|_1 \to 0$ exponentially fast. This term is sometimes called the "Markov drift" term as it arises because $\mu_t \neq \mu$. The second error term, $\delta_t(\theta_t)X_t - f_t(\theta_t)$ is "noise-like" in the sense that conditioned on the past, it is zero on expectation. As such, if the stepsize gets small enough quickly (i.e., $\sum_t \alpha_t^2 < \infty$), its effect "washes out" over time.

Thus, we find that the main affect on the evolution of θ_t comes from the update

$$\theta_{t+1} = \theta_t + \alpha_t f(\theta_t)\,. \tag{11}$$

As noted beforehand, we expect that if this iteration converges, it converges to θ_* such that v_{θ_*} is the fixed point of ΠT_π. Hence, assume now that such a fixed point exists and is unique.

Define $e_t = \theta_t - \theta_*$. Then,

$$e_{t+1} = e_t + \alpha_t f(\theta_* + e_t)\,.$$

Hence,

$$\begin{aligned}
\|e_{t+1}\|_2^2 &= \|e_t + \alpha_t f(\theta_* + e_t)\|_2^2 \\
&= \|e_t\|_2^2 + \alpha_t^2 \|f(\theta_* + e_t)\|_2^2 + 2\alpha_t(\theta_t - \theta_*)^\top f(\theta_t) \\
&\le \|e_t\|_2^2 + \alpha_t^2 \|\tilde{A}^\top \tilde{A}\|_2 \|e_t\|_2^2 + 2\alpha_t(\theta_t - \theta_*)^\top f(\theta_t)\,,
\end{aligned} \tag{12}$$

where the last inequality follows by noting that on the one hand, $f(\theta)$ is an affine linear function of θ: $f(\theta) = \tilde{A}\theta + b$ for some $b \in \mathbb{R}^d$ and $\tilde{A} \in \mathbb{R}^{d\times d}$, while on the other hand, $f(\theta_*) = 0$.[4] We will find that for small enough stepsizes α_t the length of the error e_t decreases if we can show that the last term can be upper bounded by $-c\|e_t\|_2^2$ for some positive constant c.

To study this term, it will be useful to note that by Equation (3), f satisfies the identity

$$f(\theta) = X^\top D(T_\pi v_\theta - v_\theta)\,. \tag{13}$$

For the next lemma recall that for a symmetric positive definite matrix $P \in \mathbb{R}^{d\times d}$ and $x \in \mathbb{R}^d$, $\|x\|_P^2 = x^\top P x$. We use this notation with G, which is clearly symmetric, and can also be seen to be positive definite under the condition that it is non-singular.

Lemma 5.1. *Assume that G is non-singular. Then, for any $\theta \in \mathbb{R}^d$,*

$$(\theta - \theta_*)^\top f(\theta) < (\gamma - 1)\|\theta - \theta_*\|_G^2\,.$$

Proof. Pick θ as in the statement. From Equation (13) we see that

$$(\theta - \theta_*)^\top f(\theta) = \langle v_\theta - v_{\theta_*}, T_\pi v_\theta - v_\theta \rangle_\mu\,.$$

Recall that the Cauchy-Schwarz inequality holds true: for any $u, v \in \mathbb{R}^S$,

$$\langle u, v\rangle_\mu \leq \|u\|_\mu \|v\|_\mu \,. \tag{“Cauchy-Schwarz”}$$

Hence,

$$\begin{aligned}
\langle v_\theta - v_{\theta_*}, T_\pi v_\theta - v_\theta\rangle_\mu &= \langle v_\theta - v_{\theta_*}, T_\pi v_\theta - v_{\theta_*} + v_{\theta_*} - v_\theta\rangle_\mu \\
&= \langle v_\theta - v_{\theta_*}, T_\pi v_\theta - v_{\theta_*}\rangle_\mu - \|v_{\theta_*} - v_\theta\|_\mu^2 && \text{(linearity of } v \mapsto \langle \cdot, v\rangle_\mu \text{ and } \|v\|_\mu^2 = \langle v, v\rangle_\mu\text{)} \\
&= \langle v_\theta - v_{\theta_*}, \Pi T_\pi v_\theta - \Pi T_\pi v_{\theta_*}\rangle_\mu - \|v_{\theta_*} - v_\theta\|_\mu^2 && \text{(see explanation at the end)} \\
&\leq \|v_\theta - v_{\theta_*}\|_\mu \|\Pi T_\pi v_\theta - \Pi T_\pi v_{\theta_*}\|_\mu - \|v_{\theta_*} - v_\theta\|_\mu^2 && \text{(by Cauchy-Schwarz)} \\
&\leq \gamma \|v_\theta - v_{\theta_*}\|_\mu \|v_\theta - v_{\theta_*}\|_\mu - \|v_{\theta_*} - v_\theta\|_\mu^2 && \text{(by Theorem 4.1)} \\
&= (\gamma - 1)\|v_\theta - v_{\theta_*}\|_\mu^2 \\
&= (\gamma - 1)\|\theta - \theta_*\|_G^2 \,.
\end{aligned}$$

Above, in the third equality, we used that on the one hand, by Theorem 4.1, v_{θ_*} is the fixed point of ΠT_π, which means that $v_{\theta_*} = \Pi T_\pi v_{\theta_*}$, while on the other hand, for any $\theta \in \mathbb{R}^d$ and $u \in \mathbb{R}^S$, by Equation (8),

$$\langle v_\theta, u\rangle_\mu = \langle v_\theta, \Pi u\rangle_\mu \,,$$

while $u \mapsto \langle u, v\rangle_\mu$ is linear, hence $\langle v_\theta - v_{\theta_*}, T_\pi v_\theta\rangle_\mu = \langle v_\theta - v_{\theta_*}, \Pi T_\pi v_\theta\rangle_\mu$. □

Since G is symmetric, positive definite, all its eigenvalues are real and positive. Let $\lambda_{\min}$ denote the smallest of the eigenvalues of G. Note that $\lambda_{\min} > 0$ since G is non-singular. It is known that for any $x \in \mathbb{R}^d$,

$$\|x\|_G^2 = x^\top G x \geq \lambda_{\min} \|x\|_2^2 \,.$$

Combining the result of Lemma 5.1 and the above equation with Equation (12), gives us

$$\begin{aligned}
\|e_{t+1}\|_2^2 &\leq \|e_t\|_2^2 + \alpha_t^2 \|\tilde{A}^\top \tilde{A}\|_2 \|e_t\|_2^2 + 2\alpha_t(\gamma - 1)\|e_t\|_G^2 \\
&\leq \left(1 + 2\alpha_t(\gamma - 1)\lambda_{\min} + \alpha_t^2 \|\tilde{A}^\top \tilde{A}\|_2\right) \|e_t\|_2^2 \,.
\end{aligned} \tag{14}$$

This gives rise to the following theorem:

Theorem 5.2 (Convergence of Equation (11))**.** *Assume that G is non-singular and let θ^* be the unique vector in $\mathbb{R}^d$ such that v_{θ^*} is the fixed point of ΠT_π. Consider the sequence $(\theta_t)_{t \geq 0}$ defined by Equation (11). Let $\alpha_t = \alpha$ for $t \geq 0$. Assuming that α is sufficiently small, $\theta_t \to \theta_*$ and the convergence speed is geometric.*

Proof. From Equation (14) we have $\|e_{t+1}\|_2 \leq \rho \|e_t\|_2$ where $\rho < 1$ provided that

$$1 + 2\alpha(\gamma - 1)\lambda_{\min} + \alpha^2 \|\tilde{A}^\top \tilde{A}\|_2 < 1 \,.$$

The above is equivalent to

$$\alpha(2(\gamma - 1)\lambda_{\min} + \alpha \|\tilde{A}^\top \tilde{A}\|_2) < 0 \,.$$

The largest root of the left-hand side (which is viewed as a quadratic function of α) is

$$\alpha_2 = \frac{2(1-\gamma)\lambda_{\min}}{\|\tilde{A}^\top \tilde{A}\|_2} .$$

Hence, it suffices if $0 < \alpha < \alpha_2$. □

A slightly more complicated argument can show that if $\alpha_t \to 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$, we still have $\theta_t \to \theta_*$.

5.1 From TD(0) to fitted value iteration

Using J_θ from the previous section, the mean update equation of TD(0), Equation (11), rewrites as

$$\theta_{t+1} = \theta_t - \alpha_t \nabla J_{\theta_t}(\theta_t) .$$

This can be recognized as taking a step in the direction of the negative gradient of the loss J_{θ_t}.

If we let $\theta_t^{(i+1)} = \theta_t^{(i)} - \alpha_t J_{\theta_t}(\theta_t^{(i)})$, $\theta_t^{(0)} = \theta_t$, and α_t chosen to be not too big, then $(\theta_t^{(i)})_{i \geq 0}$ can be shown to converge to the minimizer of J_{θ_t}:

$$\lim_{i\to\infty} \theta_t^{(i)} = \arg\min_{\theta} J_{\theta_t}(\theta) .$$

Thus, we can view TD(0) as an approximate, "optimistic" version of the method that updates the parameter vector via

$$\theta_{t+1} = \arg\min_{\theta'} J_{\theta_t}(\theta') . \tag{15}$$

Letting $v_t = v_{\theta_t} (= X\theta_t)$, by the definition of Π and J_θ,

$$v_{t+1} = \Pi T_\pi v_t . \tag{16}$$

By Theorem 4.1, $F = \Pi T_\pi$ is a γ-contraction with respect to $\|\cdot\|_\mu$. Hence, the iteration Equation (16) converges to the unique fixed point of F under the assumption that G is non-singular.

The algorithm given by Equation (15) is known as an instance of **fitted value iteration** (used for evaluating a policy π). The origin of the name is clear from Equation (16): Like value iteration, the algorithm iterates using T_π, except that the application of T_π is immediately followed by an application of the projection operator Π, which maps $T_\pi v_t$ back to the space $H = \{v_\theta \,:\, \theta \in \mathbb{R}^d\}$.

This can also be generalized to work with non-linear function approximation. The definition of $J_\theta(\theta')$ remains the same:

$$J_\theta(\theta') = \frac{1}{2}\|T_\pi v_\theta - v_{\theta'}\|_\mu^2 ,$$

except that now $(v_\theta)_{\theta\in\mathbb{R}^d}$ is allowed to be any parametric family of functions mapping $\mathcal{S}$ to the reals, such as neural networks. The sample-based version of this method can be written as

$$\theta_{t+1} = \arg\min_{\theta} \sum_{t=1}^{n} (R'_t + \gamma v_{\theta_t}(S'_t) - v_\theta(S_t))^2 ,$$

where (S_t, R'_t, S'_t) are such that $(R'_t, S'_t) \sim p(\cdot,\cdot|S_t, A_t)$ where $A_t \sim \pi(\cdot|S_t)$, for $t = 1, \ldots, n$.

Notes

1. We follow the convention that a Markov process with a countable state space is called a Markov chain.

2. The Markov chain whose state space is the set of natural numbers $\mathbb{N}$ and whose transition probabilities are defined via $P_{s,s+1} = 1$ for all $s \in \mathbb{N}$, does not have a stationary distribution. (Note that this Markov chain is not finite.)

3. Above we used $\nabla_{\theta'}$ to signify that the derivative whose transpose is the gradient, considered here, is with respect to θ'.

4. Here, $\|M\|_2$ is the induced 2-norm of matrix M: $\|M\|_2 = \sup_{x:\|x\|_2=1} \|Mx\|_2$. The induced 2-norm has the property that $\|Mx\|_2 \leq \|M\|_2 \|x\|_2$. This, together with Cauchy-Schwarz, which states that $|x^\top y\| \leq \|x\|_2 \|y\|_2$, gives the result.