Approximate Temporal Difference Learning is a Gradient Descent for Reversible Policies

Yann Ollivier

Abstract

In reinforcement learning, temporal difference (TD) is the most direct algorithm to learn the value function of a policy. For large or infinite state spaces, exact representations of the value function are usually not available, and it must be approximated by a function in some parametric family.

However, with *nonlinear* parametric approximations (such as neural networks), TD is not guaranteed to converge to a good approximation of the true value function within the family, and is known to diverge even in relatively simple cases. TD lacks an interpretation as a stochastic gradient descent of an error between the true and approximate value functions, which would provide such guarantees.

We prove that approximate TD is a gradient descent provided the current policy is *reversible*. This holds even with nonlinear approximations.

A policy with transition probabilities $P(s, s')$ between states is reversible if there exists a function μ over states such that $\frac{P(s,s')}{P(s',s)} = \frac{\mu(s')}{\mu(s)}$. In particular, every move can be undone with some probability. This condition is restrictive; it is satisfied, for instance, for a navigation problem in any unoriented graph.

In this case, approximate TD is exactly a gradient descent of the *Dirichlet norm*, the norm of the difference of *gradients* between the true and approximate value functions. The Dirichlet norm also controls the bias of approximate policy gradient. These results hold even with no decay factor ($\gamma = 1$) and do not rely on contractivity of the Bellman operator, thus proving stability of TD even with $\gamma = 1$ for reversible policies.

The temporal difference (TD) algorithm is a cornerstone of reinforcement learning, allowing for computation of the Bellman value function of a given policy [SB98]. However, with large or continuous search spaces, maintaining the exact value function at each state is unfeasible, and parametric approximations of the value function are used instead [SB98, §8].

With such parametric approximations, TD is not guaranteed to converge to the best approximation of the true value function within the family, or even, to converge at all [TVR97, §X]. This is in great part because the TD algorithm lacks an interpretation as a stochastic gradient descent of an error between the true and approximate value functions.

For *linear* families of approximating functions, TD is known to converge to some fixed point [TVR97]; this fixed point is related, but generally not identical, to the best approximation in the family. For nonlinear approximations, TD is known to diverge even in relatively simple cases. Current popular families using neural networks are nonlinear.

As a theoretical study of nonlinear value function approximation, [MSB$^+$09] introduces an algorithm more complex than TD, involving second derivatives of the approximating family. This algorithm has an interpretation as a gradient descent of an objective function J. J is built so that the global minimum of J is also a fixed point of TD; however, the algorithm may also converge to a local minimum of J with unclear significance. Moreover this does not address the interpretation of fixed points of TD in the first place.

Here we consider the unmodified approximate TD algorithm, with any class of approximating functions, linear or not. We prove that approximate TD coincides with a gradient descent of the *Dirichlet norm* of the error between the true and approximate value functions (Theorem 1), provided the current policy is *reversible*.

Reversibility (see Section 1) is a common assumption in the mathematical treatment of Markov chains, because of its convenience. It implies that any allowed transition between states can also occur in reverse with some probability. It is satisfied, for instance, by the random walk on unoriented graphs, or by Brownian motion and other stochastic processes.

The Dirichlet norm is used in the treatment of the convergence of Markov chains [DSC96, LPW09], and is directly related to the spectral gap of the random walk operator. This norm is given in a simple way by the transition probabilities of the current policy (Eq. 11). Its natural appearance in approximate TD is perhaps remarkable.

Therefore, approximate TD learning will minimize the approximation error in Dirichlet norm, for reversible policies. Interestingly, this minimization also directly controls the bias of approximate policy gradient, which also involves the Dirichlet norm (Proposition 4).

However, in a reinforcement learning setting, the reversibility assumption is quite restrictive. First, it implies that any move can be undone with some probability. Second, reversibility depends both on the policy and the environment (via Eq. 2); in general, reversibility cannot be checked knowing the policy alone. An exception to this are *navigation-type* problems, in which the policy consists in directly choosing the next state among a set of possible states (e.g., exploring an undirected graph). For such problems, it is easy to check reversibility, and to keep the policy reversible at all times, e.g., by using a Gibbs policy with respect to some energy function on state space (see Section 4).

Thus, although we have stated each result under general mathematical assumptions, the results here chiefly make sense in navigation-type problems, in which the agent directly selects the next state among a set of neighbors, and any move can be reversed.

Acknowledgments. I would like to thank Léon Bottou, Alessandro Lazaric, Corentin Tallec and Nicolas Usunier for pointers to references and for suggestions on the text.

1 Notation and Markov Chain Background

Markov decision processes. We mostly borrow notation from [MSB+09]. Consider a finite[1] Markov decision process (MDP) and a policy π for this MDP. Let $\pi(s,a)$ be the probability to select action a when in state s. Let $P_{\text{env}}((s,a),s')$ the probability that the environment jumps to s' after that. Let $r_1,\ldots,r_t$ be the sequence of rewards of this MDP: r_t is the reward incurred while arriving in state s_t, a random variable depending on a_{t-1} and s_{t-1}.

Given an initial state s_0, denote $\mathbb{E}_{\pi,s_0}$ the expectation under a random sequence $(a_0,s_1,a_1,\ldots)$ of actions and states resulting from π and P_{env}, defined inductively by $a_t \sim \pi(s_t,\cdot)$ and $s_{t+1} \sim P_{\text{env}}((s_t,a_t),\cdot)$.

The *value function* of policy π in state s, with *decay parameter* $\gamma < 1$, is

$$V(s) := \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{E}_{\pi,s}[r_t] \tag{1}$$

Define the transition probability matrix P on states, that amounts to first selecting an action according to π, then letting the environment select the next state [MSB+09]:

$$P(s,s') := \sum_a \pi(s,a) P_{\text{env}}((s,a),s') \tag{2}$$

The value function for policy π satisfies the Bellman equation using transition probabilities P,

$$V = R + \gamma P V \tag{3}$$

where P and V are seen as a matrix and vector, and

$$R(s) := \mathbb{E}_{\pi,s}[r_1] \tag{4}$$

is the average instantaneous reward of the policy in a given state.

[1] The arguments presented here do not crucially rely on finiteness: algebraically the results would hold for a countable or continuous state space as well, as long as all sums and expectations are well-defined. We consider the finite case to avoid measurability issues.

For $\gamma = 1$ the value function is usually infinite. We use the *relative* value function [SB98, §6.7], also known as *bias* [Ber12, §5.1.1], denoted U. Assuming that the current policy has a unique stationary distribution μ over states, the relative value function is obtained by centering rewards:

$$U(s) := \sum_{t=1}^{\infty} \mathbb{E}_{\pi,s}[r_t - \mathbb{E}_\mu R] \tag{5}$$

where $\mathbb{E}_\mu R = \sum \mu(s)R(s)$ is the average reward under the stationary distribution μ. (Assuming ergodicity of P, this expectation is finite in a finite MDP [Ber12, §5.1.1], though without the expectation the sum usually diverges as noise accumulates.) The relative value function satisfies the Bellman equation with $\gamma = 1$ and centered rewards [Ber12, Prop. 5.1.9]

$$U = (R - \mathbb{E}_\mu R) + PU \tag{6}$$

Approximate TD. Let V_θ be an approximation to the true function V, belonging to some family of functions smoothly parameterized by θ.

Given a transition $s \to s'$ with reward r, the gap in the Bellman equation at s is $r + \gamma V_\theta(s') - V_\theta(s)$. For the true V function, this gap is 0 on average (on average, because given s, the state s' and the reward are random). Approximate TD (e.g. [SB98, §8.2] with $\lambda = 0$) performs an update on $V_\theta(s)$ to reduce the gap,

$$\theta \leftarrow \theta + \alpha\,\Delta\theta \tag{7}$$

where α is a learning rate and $\Delta\theta$ is the update

$$\Delta\theta(s, s', r) := (r + \gamma V_\theta(s') - V_\theta(s))\,\partial_\theta V_\theta(s) \tag{8}$$

This gradient step has the effect of moving $V_\theta(s)$ closer to the *current* value of $r + \gamma V_\theta(s')$, ignoring the fact that $V_\theta(s')$ will change as well.

Reversibility of Markov chains. A Markov chain defined by the transition matrix P is *reversible* [LPW09, §1.6] if there exists a nonzero function μ on states such that

$$\mu(s)P(s, s') = \mu(s')P(s', s) \quad \forall s, s' \tag{9}$$

When nonzero this rewrites as $P(s, s')/P(s', s) = \mu(s')/\mu(s)$: the ratio between the probability of a transition and the reverse transition must be equal to a ratio of a function of the target states. In particular, any states s and s' with nonzero μ must satisfy $P(s, s') > 0 \Leftrightarrow P(s', s) > 0$.

For instance, the simple random walk in any *unoriented* graph is reversible with $\mu(s) = \deg(s)$ [LPW09, §1.6].

When P is reversible with respect to μ, then μ (once rescaled) is a stationary distribution of P [LPW09, Prop. 1.19]. Indeed, the condition above describes *detailed balance*: if starting from distribution μ, the flow of mass from s to s' is equal to that from s' to s, so that every exchange is balanced and μ is stationary.

Therefore, reversibility of a Markov chain is usually expressed directly with respect to its stationary distribution μ.

On any unoriented graph, the Metropolis–Hastings construction provides reversible random walks with arbitrary stationary distributions (see Section 4). Thus, for navigation problems on states spaces with reversible moves, it would be easy to keep the policy reversible.

By abuse of language, in a reinforcement learning context within a fixed environment, we will call a policy *reversible* if the Markov chain P defined by this policy in that environment via (2) is reversible.

The Dirichlet norm for Markov chains. Given a function f on the state space, define its square norm under the stationary distribution μ, and the associated bilinear form, as

$$\|f\|_\mu^2 := \sum_s \mu(s) f(s)^2, \qquad \langle f, g\rangle_\mu := \sum_s \mu(s) f(s) g(s) \tag{10}$$

The weighting by $\mu(s)$ is perhaps best interpreted as an average over a long trajectory sampled from the policy.

The Markov chain is reversible with respect to μ if and only if P is self-adjoint for this bilinear form, namely, if and only if $\langle Pf, g\rangle_\mu = \langle f, Pg\rangle_\mu$, where P acts on a function f over states by viewing P as a matrix and f as a vector. This is a direct consequence of (9).

We also define the *Dirichlet norm* depending on the transition matrix P:

$$\|f\|_{\mathrm{Dir}}^2 := \frac{1}{2} \sum_{s,s'} \mu(s) P(s,s') (f(s') - f(s))^2 \tag{11}$$

where μ is the invariant distribution on states resulting from P. This quadratic form is actually a seminorm, since constant functions have norm 0: adding a constant to f does not change $\|f\|_{\mathrm{Dir}}$. If P is irreducible then constant functions are the only such functions: if $\|f_1 - f_2\|_{\mathrm{Dir}} = 0$ then f_1 and f_2 are equal up to an additive constant. This justifies the name *norm* if quotienting by constant functions.

$\|f\|_{\mathrm{Dir}}^2$ is often called the *Dirichlet form* in the Markov chain literature [DSC96, LPW09]. It is a discrete Markov chain analogue of the gradient norm $\int \|\nabla f\|^2$ of a continuous function (the classical "Dirichlet form"): indeed, for f a smooth function with compact support in $\mathbb{R}^d$, and P the nearest-neighbor random walk on an ε-grid in $\mathbb{R}^d$, with $\varepsilon \ll 1$, at any point x in the grid one has

$$\sum_{x'} P(x,x')(f(x') - f(x))^2 = \frac{\varepsilon^2}{d} \|\nabla f(x)\|^2 + O(\varepsilon^3) \tag{12}$$

by a direct Taylor expansion, and therefore

$$\|f\|_{\mathrm{Dir}}^2 = \frac{\varepsilon^2}{2d}\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 \,\mathrm{d}x + O(\varepsilon^3) \tag{13}$$

By elementary computations, the Dirichlet norm satisfies [DSC96, LPW09]

$$\|f\|_{\mathrm{Dir}}^2 = \langle (\mathrm{Id} - P)f, f\rangle_\mu \tag{14}$$

These two norms control one another up to centering: for any f,

$$\beta\,\|f - \mathbb{E}_\mu f\|_\mu^2 \leqslant \|f\|_{\mathrm{Dir}}^2 \leqslant \|f - \mathbb{E}_\mu f\|_\mu^2 \leqslant \|f\|_\mu^2 \tag{15}$$

with β the *spectral gap* of the random walk [DSC96]. In practice β may be quite small: e.g., for the simple random walk on a cycle of length n, one has $\beta \approx 1/n^2$. Therefore $\|\cdot\|_{\mathrm{Dir}}$ can be significantly smaller than $\|\cdot\|_\mu$.

2 Approximate TD for Reversible Policies

We claim that *if P is reversible with respect to its stationary distribution μ*, then approximate TD learning with a class of functions V_θ, tries to best approximate the true function V by gradient descent. The quality of the approximation is defined via a mixed norm of $V_\theta - V$,

$$\gamma\,\|V_\theta - V\|_{\mathrm{Dir}}^2 + (1-\gamma)\,\|V_\theta - V\|_\mu^2 \tag{16}$$

where V is the true Bellman function associated with policy P.

For γ close to 1, the Dirichlet norm $\|V_\theta - V\|_{\mathrm{Dir}}$ dominates, while for small γ the μ-norm dominates. (For $\gamma = 0$ the V-function is equal to the expected instantaneous reward.)

Thus, assuming reversibility, approximate TD will usually converge to a local minimum of this mixed norm of $V_\theta - V$. This is independent of the family of parametric approximations for V. For the particular case of a linear family over θ, the mixed norm is quadratic in θ, therefore convergence will be to a global minimum of the mixed norm. The equivalence of the norms (15) can be used to transfer the minimization property to either $\|\cdot\|_\mu$ or $\|\cdot\|_{\mathrm{Dir}}$ up to factors β.

Theorem 1. *Consider a policy in some finite MDP. Assume the policy is reversible, with stationary distribution μ.*

Let V be the value function of the policy with decay factor $0 \leqslant \gamma < 1$. Let $(V_\theta(s))_\theta$ be a family of functions on the state space, smoothly parameterized by θ.

Then, on average over the stationary distribution μ, the step $\Delta\theta(s, s', r)$ made by approximate TD (8) is equal to a gradient descent of a mixed norm of $V_\theta - V$,

$$\mathbb{E}_{s\sim\mu}\,\Delta\theta(s, s', r) = -\frac{1}{2}\partial_\theta\left(\gamma\,\|V_\theta - V\|_{\mathrm{Dir}}^2 + (1-\gamma)\,\|V_\theta - V\|_\mu^2\right) \tag{17}$$

where s' and r are the (random) next state and reward from state s.

The theorem is in expectation over states s from the stationary distribution. Averaging over a long enough trajectory, with small enough learning rates, will approximate this expectation. [2]

At the core of the proof, TD only takes into account cross-terms between $\partial_\theta V_\theta$ at the current state and the value function at the next state, while the gradient of the error between V_θ and V also comprises cross-terms between $\partial_\theta V_\theta$ at the next state and the value function at the current state. In the reversible case, the statistics of transitions $s \to s'$ and $s' \to s$ are identical in the stationary regime, hence TD is indeed a gradient of the error.

Proof.
The expected TD step in the stationary regime is

$$\begin{aligned}
\mathbb{E}_{s\sim\mu}\,\Delta\theta(s,s',r) &= \sum_s \mu(s)\partial_\theta V_\theta(s)\,\mathbb{E}_{s'|s}\left[r+\gamma V_\theta(s')-V_\theta(s)\right] && (18)\\
&= \sum_s \mu(s)\partial_\theta V_\theta(s)\left(R+\gamma(PV_\theta)(s)-V_\theta(s)\right) && (19)\\
&= \langle \partial_\theta V_\theta, R+\gamma PV_\theta - V_\theta\rangle_\mu && (20)
\end{aligned}$$

namely, the expected TD step is the dot product between the Bellman gap V_θ, and the direction of change $\partial_\theta V_\theta$ that can be realized within the parametric family. (In the linear case, this reduces to, e.g., Lemma 8 in [TVR97], with $\partial_\theta V_\theta = \Phi$.)

Define the difference between the approximated and true V functions:

$$f_\theta := V_\theta - V \tag{21}$$

we want to prove that the expected TD step is the gradient of the mixed norm of f_θ.

Since V satisfies the Bellman equation $R+\gamma PV - V = 0$ one has

$$R+\gamma PV_\theta - V_\theta = \gamma P f_\theta - f_\theta \tag{22}$$

and moreover $\partial_\theta V_\theta = \partial_\theta f_\theta$ as V does not depend on θ. Therefore, (20) rewrites as

$$\begin{aligned}
\mathbb{E}_{s\sim\mu}\,\Delta\theta(s,s',r) &= \langle \partial_\theta f_\theta, (\gamma P - \mathrm{Id})f_\theta\rangle_\mu && (23)\\
&= -\gamma\,\langle \partial_\theta f_\theta, (\mathrm{Id}-P)f_\theta\rangle_\mu - (1-\gamma)\langle \partial_\theta f_\theta, f_\theta\rangle_\mu && (24)
\end{aligned}$$

Now the last term is the gradient of the μ-norm:

$$\langle \partial_\theta f_\theta, f_\theta\rangle_\mu = \frac{1}{2}\partial_\theta \langle f_\theta, f_\theta\rangle_\mu = \frac{1}{2}\partial_\theta \left\|f_\theta\right\|_\mu^2 \tag{25}$$

[2] TD is a stochastic update whose noise depends on s, so that the noise is Markov instead of iid. The general theory of stochastic algorithms with Markov noise from [BMP90] is used in [TVR97] to offer a full treatment of TD for linear approximations V_θ.

Likewise, the first term is related to the norm $\|\cdot\|^2_{\mathrm{Dir}}$ thanks to (14):

$$\|f_\theta\|^2_{\mathrm{Dir}} = \langle f_\theta, (\mathrm{Id} - P) f_\theta\rangle_\mu \tag{26}$$

hence

$$\partial_\theta \|f_\theta\|^2_{\mathrm{Dir}} = \langle \partial_\theta f_\theta, (\mathrm{Id} - P) f_\theta\rangle_\mu + \langle f_\theta, (\mathrm{Id} - P)\partial_\theta f_\theta\rangle_\mu \tag{27}$$

as $\mathrm{Id} - P$ is a linear operator that does not depend on θ.

But the policy is reversible with respect to μ if and only if P is self-adjoint with respect to $\langle\cdot,\cdot\rangle_\mu$. In that case,

$$\langle f_\theta, (\mathrm{Id} - P)\partial_\theta f_\theta\rangle_\mu = \langle (\mathrm{Id} - P) f_\theta, \partial_\theta f_\theta\rangle_\mu \tag{28}$$

and therefore

$$\partial_\theta \|f_\theta\|^2_{\mathrm{Dir}} = 2\langle \partial_\theta f_\theta, (\mathrm{Id} - P) f_\theta\rangle_\mu \tag{29}$$

Collecting (25) and (29) into (24), we find

$$\mathbb{E}_{s\sim\mu}\, \Delta\theta(s, s', r) = -\frac{1}{2}\gamma\, \partial_\theta \|f_\theta\|^2_{\mathrm{Dir}} - \frac{1}{2}(1-\gamma)\, \partial_\theta \|f_\theta\|^2_\mu \tag{30}$$

as needed. □

In the general, non-reversible case, the gradient descent of $\|V_\theta - V\|^2_{\mathrm{Dir}}$ differs from TD by

$$2\sum_{s,s'} \mu(s)P(s,s')\partial V_\theta(s') \left(V_\theta(s') - V_\theta(s) - V(s') + V(s)\right) \tag{31}$$

which we cannot compute without knowing V. At least we would have to know how to estimate $\mathbb{E}_{s|s'} V(s) - V(s')$ given s'. That is, we would need to be able to sample backward transitions leading to s', and to evaluate the reward along these transitions. This is similar to attempting to take the gradient of the squared Bellman error [SB98, §8.5].

We now turn to the case $\gamma = 1$. The relative value function U can be approximated by using approximate TD with *centered* rewards, namely, by removing the stationary expected reward at each step [SB98, §6.7]. In practice the expected reward is usually unknown and must be approximated by averaging over the past.

Theorem 2. *Consider a policy in some finite MDP. Assume the policy is reversible, with stationary distribution μ.*

Let U be the relative value function of the policy (with decay factor $\gamma = 1$). Let $(U_\theta(s))_\theta$ be a family of functions on the state space, smoothly parameterized by θ. Let $\Delta\theta(s, s', r)$ be the step made by centered approximate TD during a transition $s \to s'$ with reward r, namely

$$\Delta\theta(s, s', r) := \left(r - \mathbb{E}_\mu R + U_\theta(s') - U_\theta(s)\right) \partial_\theta U_\theta(s) \tag{32}$$

Then, on average over the stationary distribution μ, the step made by centered approximate TD is equal to a gradient descent of the Dirichlet norm of $U_\theta - U$,

$$\mathbb{E}_{s\sim\mu}\,\Delta\theta(s,s',r) = -\frac{1}{2}\,\partial_\theta\,\|U_\theta - U\|^2_{\mathrm{Dir}} \tag{33}$$

where s' and r are the (random) next state and reward from state s.

Proof.
The proof is strictly identical, replacing V with U, discarding all $(1-\gamma)$ terms, and using that U satisfies the centered Bellman equation (6). In particular, the Dirichlet norm is insensitive to adding constants, so the centering of rewards does not affect the result. □

Advantage function over states, and Dirichlet norm. Take $\gamma = 1$. Given a transition $s \to s'$, define the advantage of s' at s to be

$$A(s'|s) := \mathbb{E}[r(s,s')] + U(s') - U(s) \tag{34}$$

and likewise the approximate advantage $A_\theta(s'|s) := \mathbb{E}[r(s,s')] + U_\theta(s') - U_\theta(s)$. This is the "state advantage function", defined as a function of the next state s', as opposed to the usual advantage function which is defined on actions. Once more, this is relevant mostly in a navigation setting where actions directly correspond to choosing the next state.

Then the Dirichlet norm of $U_\theta - U$ is the average square error of the advantage function:

$$\mathbb{E}_{s\sim\mu}\mathbb{E}_{s'\sim P(s,s')}(A(s'|s) - A_\theta(s'|s))^2 = 2\,\|U - U_\theta\|^2_{\mathrm{Dir}} \tag{35}$$

by direct substitution. Therefore, Theorem 2 can be restated using this advantage function.

Corollary 3. *For $\gamma = 1$ and for reversible policies, centered approximate TD is a gradient descent of the average square error $\mathbb{E}_{s\sim\mu}\mathbb{E}_{s'\sim P(s,s')}(A(s'|s) - A_\theta(s'|s))^2$ of the state advantage function.*

However, for $\gamma < 1$ this correspondence breaks down. Indeed, defining the state advantage function for $\gamma < 1$ as

$$A(s'|s) := \mathbb{E}[r(s,s')] + \gamma V(s') - V(s) \tag{36}$$

and likewise for A_θ, one checks that

$$\mathbb{E}_{s\sim\mu}\mathbb{E}_{s'\sim P(s,s')}(A(s'|s) - A_\theta(s'|s))^2 = 2\gamma\,\|V - V_\theta\|^2_{\mathrm{Dir}} + (1-\gamma)^2\,\|V - V_\theta\|^2_\mu \tag{37}$$

which is not quite the mixed norm minimized by TD: the weights between the two norms are different.

3 The Dirichlet Norm and Policy Gradient Bias

We have proved that with reversible policies, TD approximates the value function in the Dirichlet norm. This clarifies the behavior of TD for policy evaluation, but does this help with policy improvement?

Classical results state that if an approximate value function is ε-close to the true value function (in sup norm), then greedy policies based on the approximate value function will have cumulated rewards that are $2\varepsilon/(1-\gamma)$-close to the optimal cumulated rewards [Ber12, Prop 2.3.3].

The Dirichlet norm, on the other hand, controls how close policy gradient based on the true or approximate value functions are to each other: this is Proposition 4 below. Interestingly, this directly holds with $\gamma = 1$, without factors $1/(1-\gamma)$.

In a non-episodic setting, policy gradient is defined as the gradient of the expected reward under the stationary distribution of the policy [Ber12, §7.4]: the goal is to maximize the average reward collected along an infinitely long trajectory of this policy.

So let π_φ be a policy smoothly parameterized by φ. Let μ_φ be the stationary distribution of π_φ. Here we do *not* assume that policies are reversible.

The expected reward of the policy with parameter φ is

$$\mathcal{R}(\varphi) := \sum_s \mu_\varphi(s) R(s) \tag{38}$$

with $R(s)$ the expected instantaneous reward in state s (which itself depends on φ via the expectation in (4)). The direction of the policy gradient update is $\partial_\varphi \mathcal{R}(\varphi)$.

The classical policy gradient theorem [Ber12, §7.4.1] provides a way to compute this gradient: it is an expectation under the stationary distribution, of the correlation between expected rewards and action probabilities. [3] The direction of the gradient can be expressed as [Ber12, Eq. (7.120)] [4]

$$\begin{aligned}\Delta\varphi &:= \partial_\varphi \mathcal{R}(\varphi)\\ &= \mathbb{E}_{s\sim\mu_\varphi,\, a\sim\pi_\varphi(s,a),\, s'\sim P_{\mathrm{env}}((s,a),s')}\left[\left(r(s,a,s') + U(s')\right)\partial_\varphi \ln \pi_\varphi(s,a)\right]\end{aligned} \tag{39}$$

with U the relative value function of the current policy, and $r(s,a,s')$ the random reward incurred during the transition $s \to s'$.

The policy gradient $\Delta\varphi$ is an expectation over transitions (s,a,s') from the current policy. As such, an algorithm averaging over long trajectories

[3] This assumes the environment is independent from the parameter φ used by the agent.

[4] Eq. (7.120) in [Ber12] uses centered rewards in the definition of $\tilde{Q}$. This is indifferent: since $\sum_a \partial_\varphi \ln \pi_\varphi(s,a) = 0$, any constant or baseline can be subtracted.

from this policy would be a stochastic gradient descent with expected step $\Delta\varphi$. (See also the note after Theorem 1.)

Using an approximation of U in (39) would result in a bias; we show that this bias is controlled by the Dirichlet norm of the approximation of U.

Proposition 4. *Let $\hat{U}$ be any approximation of the relative value function U of the policy π_φ (undiscounted, $\gamma = 1$). Let $\widehat{\Delta\varphi}$ be the approximate policy gradient computed from $\hat{U}$, namely*

$$\widehat{\Delta\varphi} := \mathbb{E}_{s\sim\mu_\varphi,\, a\sim\pi_\varphi(s,a),\, s'\sim P_{\text{env}}((s,a),s')} \left[\left(r(s,a,s') + \hat{U}(s')\right) \partial_\varphi \ln \pi_\varphi(s,a)\right] \tag{40}$$

Then the bias of this approximate policy gradient is at most

$$\left\|\widehat{\Delta\varphi} - \Delta\varphi\right\|^2 \leqslant 2\left\|U - \hat{U}\right\|^2_{\text{Dir}} \cdot \left(\mathbb{E}_{s\sim\mu_\varphi} \mathbb{E}_{a\sim\pi_\varphi(s,a)} \left\|\partial_\varphi \ln \pi_\varphi(s,a)\right\|^2\right) \tag{41}$$

As a consequence, if U tends to $\hat{U}$ in Dirichlet norm then the bias tends to 0. Of course this is the bias over a single step of policy gradient. [SRB11] contains a full study of the asymptotic bias produced by a bias at each step of a stochastic gradient descent, under convexity assumptions (which would hold close to a nondegenerate local minimum in typical cases); in particular, under strong convexity, a bounded bias at each step of a gradient descent only produces a bounded deviation from the true trajectory [SRB11, Prop. 3].

Since $\|\cdot\|_{\text{Dir}} \leqslant \|\cdot\|_\mu$, the inequality also holds with $\left\|U - \hat{U}\right\|_\mu$, but is less sharp (for instance, $\|\cdot\|_{\text{Dir}}$ is insensitive to adding a constant to $\hat{U}$), sometimes much less so depending on the spectral gap β in (15).

The last factor, $\mathbb{E}_{s\sim\mu_\varphi}\mathbb{E}_{a\sim\pi_\varphi(s,a)} \left\|\partial_\varphi \ln \pi_\varphi(s,a)\right\|^2$, does not depend on the way the value function is approximated: it depends only on the way policies are parameterized. It is equal to the trace of the Fisher information matrix of the policy $\pi_\varphi(s,a)$ with respect to φ. Thus, there is a clear contribution from value function approximation, and another from the geometry of the space of policies.

Proof.

The proof is essentially the Cauchy–Schwarz inequality after subtracting a suitable baseline.

For short, denote

$$\xi(s,a,s') := \mu_\varphi(s)\pi_\varphi(s,a)P_{\text{env}}((s,a),s') \tag{42}$$

the stationary distribution over transitions (s,a,s') when using policy π_φ.

Gradients of log-probabilities have expectation 0, so for any state s,

$$\mathbb{E}_{a\sim\pi_\varphi(s,a)}\, \partial_\varphi \ln \pi_\varphi(s,a) = 0 \tag{43}$$

therefore, in the policy gradient formula (39) we can subtract any baseline depending on s, as is often done in practice:

$$\Delta\varphi = \mathbb{E}_{(s,a,s')\sim\xi}\left[\left(r(s,a,s') + U(s') - U(s)\right)\partial_\varphi \ln \pi_\varphi(s,a)\right] \tag{44}$$

and likewise

$$\widehat{\Delta\varphi} = \mathbb{E}_{(s,a,s')\sim\xi}\left[\left(r(s,a,s') + \hat{U}(s') - \hat{U}(s)\right)\partial_\varphi \ln \pi_\varphi(s,a)\right] \tag{45}$$

therefore

$$\widehat{\Delta\varphi} - \Delta\varphi = \mathbb{E}_{(s,a,s')\sim\xi}\left[\left(\hat{U}(s') - U(s') - \hat{U}(s) + U(s)\right)\partial_\varphi \ln \pi_\varphi(s,a)\right] \tag{46}$$

so by the Cauchy–Schwarz inequality

$$\left\|\widehat{\Delta\varphi} - \Delta\varphi\right\|^2 \leqslant \left(\mathbb{E}_{(s,a,s')\sim\xi}\left(\hat{U}(s') - U(s') - \hat{U}(s) + U(s)\right)^2\right)\cdot \left(\mathbb{E}_{(s,a,s')\sim\xi}\left\|\partial_\varphi \ln \pi_\varphi(s,a)\right\|^2\right) \tag{47}$$

Now marginalizing $\xi(s,a,s')$ over a yields $\mu_\varphi(s)P_\varphi(s,s')$ by definition (2). So by definition of the Dirichlet norm (11), the first factor above is exactly $2\left\|\hat{U} - U\right\|^2_{\mathrm{Dir}}$. □

Note that the Dirichlet norm itself depends on the parameter φ, via P and μ.

4 Discussion and Conclusion

Advantage function, and Dirichlet norm versus L^2 norm. Converging to the true value function in L^2 norm emphasizes getting the correct value at each state. On the other hand, converging to the true value function in Dirichlet norm emphasizes getting the correct *differences of values* between consecutive states: this is clear from the definition (11). Getting these differences right amounts to being able to compare the values of states. Proposition 4 formalizes this intuition: the smaller the error in Dirichlet norm, the smaller the bias in policy gradient. This is also directly related to the advantage function (Eq. 35) for $\gamma = 1$: for reversible policies and $\gamma = 1$, TD is just a gradient descent of the L^2 error of the state advantage function (34). Here, advantages are computed on next states s', rather than on actions as is more common; so once more this is mostly relevant for navigation problems or when a good model of the environment is available.

On convergence speed when $\gamma \to 1$. In Theorem 1, the properties of TD do not deteriorate when $\gamma \to 1$: the Dirichlet norm $\|V_\theta - V\|_{\mathrm{Dir}}$ will decrease at a rate that does not depend on γ. More precisely, TD with step size η is a gradient descent of the mixed norm with step size $\eta/2$. On the other hand, the convergence proof from [TVR97] in the linear case provides smaller and smaller learning rates when $\gamma \to 1$: the rate of decrease of the error $\theta - \theta^*$ is given by Lemma 9 from [TVR97], which contains a $(\gamma - 1)$ factor (though this can be improved using the spectral gap of the Markov chain).

Optimizing among reversible policies. The reversibility constraint on the policy is obviously a major restriction of these results.

Still, the space of reversible policies is quite large. For instance, for any positive function f on any unoriented graph, the celebrated Metropolis–Hastings construction [LPW09, §3.2.2] provides a reversible random walk on the edges of the graph, whose stationary distribution is proportional to f. The probability to jump from s to s' is set to

$$P_f(s, s') = \min\left(\frac{1}{\deg(s)}, \frac{f(s')}{f(s)\deg(s')}\right) \tag{48}$$

for any adjacent states $s \neq s'$ in the graph. Then $\mu(s) := \frac{f(s)}{\sum_{s'} f(s')}$ is a reversible stationary distribution of P_f. More generally, if P_0 is any "default" Markov chain, then the Markov chain $P_f(s, s') = \min\left(P_0(s, s'), \frac{f(s')P_0(s', s)}{f(s)}\right)$ for $s \neq s'$, is reversible with respect to this same μ.

Thus, in some cases it would be possible to explicitly keep the policy in a space of reversible policies. In particular, any navigation problem where the agent directly selects the next state among a set of neighbors of the current state, defines an unoriented graph. For such problems, policies targeting any stationary distribution f over states can be obtained by parameterizing a family of positive functions f over the state space in any convenient way, and setting the family of policies to the Metropolis–Hastings Markov chain P_f for f in this family. A natural candidate would be Gibbs distributions of the form $f(s) = \exp(\beta V_\theta(s))$ where V_θ is the family used to approximate the value function: for large β this targets high-value states.

Conclusion. We have proved that the unmodified approximate TD algorithm is exactly a gradient descent of the Dirichlet norm of the error between the true and approximate value functions, provided the policy is reversible. The Dirichlet norm also controls the bias of approximate policy gradient and the L^2 error on the advantage function over states, even for non-reversible policies. However, the reversibility condition is restrictive: only for navigation problems can one easily maintain the policy within a set of reversible policies.

Thus, at least for navigation problems, the Dirichlet norm provides a coherent theoretical picture of what approximate TD does.

References

[Ber12] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, 4th edition, 2012.

[BMP90] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

[DSC96] Persi Diaconis and Laurent Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[MSB$^+$09] Hamid R Maei, Csaba Szepesvári, Shalabh Bhatnagar, Doina Precup, David Silver, and Richard S. Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in Neural Information Processing Systems*, pages 1204–1212, 2009.

[SB98] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. MIT Press, 1998.

[SRB11] Mark Schmidt, Nicolas L. Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

[TVR97] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.