University of Alberta

The Theoretical Foundation for Incremental Least-Squares Temporal Difference Learning

by

Martin Zinkevich

Technical Report TR 06-00
November 2006
Revised November 2006

DEPARTMENT OF COMPUTING SCIENCE
University of Alberta
Edmonton, Alberta, Canada

Abstract

In this paper we present a mathematical foundation for Incremental Least-Squares Temporal Difference Learning (iLSTD) for policy evaluation in reinforcement learning with linear function approximation. iLSTD is an incremental method for achieving results similar to LSTD, the data-efficient, least-squares version of temporal difference learning, without incurring the full cost of the LSTD computation. Here, we give a technical foundation for the asymptotic properties of iLSTD.

1 Introduction

This paper presents an abstract mathematical result which is fundamental to proving that iLSTD works correctly. This proof is based heavily on a theoretical result in [1]. However, we diverge from their proof in some details, for they are considering coefficients (C_t and d_t) which always have a high variance but converge quickly in expectation, whereas we are considering coefficients which have a very low variance but converge slowly.

2 Preliminaries

We begin with some standard definitions. Given a vector $x \in \mathbf{R}^n$, $\|x\| = \sum_{i=1}^n (x_i)^2$. Given any matrix $B \in \mathbf{R}^{n\times n}$, define:

$$\|B\| = \max_{x\in\mathbf{R}^n, \|x\|=1} \|Bx\| \tag{1}$$

Azuma's Inequality states that, for any $t > 0$, for any $c_1 \dots c_n$, for any sequence of random variables $X_1, \dots, X_n$, where $X_i \in [-c_i, c_i]$ which is a martingale difference sequence (for any $i \in 0 \dots n-1$, $x_1, \dots, x_i$, $\mathbf{E}[X_{i+1}|X_1 \dots X_i] = 0$):

$$\Pr[X_n \geq t] \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right) \tag{2}$$

This can also be applied to a supermartingale, where $\mathbf{E}[X_{i+1}|X_1 \dots X_i] \leq 0$. This is a very effective technique for bounding how far one will wander in a random walk.

3 Theoretical Results

In this section, we prove that the algorithm converges. We follow [1] very closely. However, we diverge from their proof in some details, for they are considering coefficients (C_t and d_t) which always have a high variance but converge quickly in expectation, whereas we are considering coefficients which have a very low variance but converge slowly.

We consider the theoretical model, where, for all t, $y_t \in \mathbf{R}^n$,$d_t \in \mathbf{R}^n$, $R_t, C_t \in \mathbf{R}^{n\times n}$, $\beta_t \in \mathbf{R}$, and:

$$y_{t+1} = y_t + \beta_t(R_t)(C_t y_t + d_t) \tag{3}$$

Define F_t to be the state of the algorithm before R_t is selected on round t (which is after the selection of C_t and d_t). It is the case that C_t and d_t are sequences of random variables.

Assumption 1 *In order to prove this converges, we assume that there is a C^*, d^*, v, $\mu > 0$, and M such that:*

1. *C^* is negative definite,*
2. *C_t converges to C^* with probability 1,*
3. *d_t converges to d^* with probability 1,*
4. *$\mathbf{E}[R_t|F_t] = I$, and $\|R_t\| \leq M$,*
5. *$\lim_{T\to\infty} \sum_{t=1}^{T} \beta_t = \infty$, and*
6. *$\beta_t < vt^{-\mu}$.*

Before continuing with the main proof, we give a result that shows how, without the randomness, convergence is exponentially fast to C_t and d_t.

Theorem 1 *(Theorem 4 in the paper) If C_t is negative definite, for some β dependent upon C_t, if $R_t = I$, then there exists an $\zeta \in (0,1)$ such that for all y_t, if $y_{t+1} = y_t + \beta(C_t y_t + d_t)$, then $\left\|y_{t+1} + (C_t)^{-1}d_t\right\| < \zeta \left\|y_t + (C_t)^{-1}d_t\right\|$.*

Proof: Define $w_t = y_t + (C_t)^{-1}d_t$, and w_{t+1} similarly. What we must prove is that, for some $\zeta < 1$, $\|w_{t+1}\| \leq \zeta \|w_t\|$.

Observe that:

$$y_{t+1} = w_t - (C_t)^{-1}d_t + \beta(C_t(w_t - (C_t)^{-1}d_t) + d_t) \tag{4}$$

$$y_{t+1} + (C_t)^{-1}d_t = w_t + \beta(C_t w_t) \tag{5}$$

$$w_{t+1} = w_t + \beta C_t w_t \tag{6}$$

Now, we consider the square of both sides:

$$\|w_{t+1}\|^2 = (w_t + \beta C_t w_t)^2 \tag{7}$$

$$\|w_{t+1}\|^2 = \|w_t\|^2 + 2\beta w_t \cdot (C_t w_t) + \beta^2 \|C_t w_t\|^2 \tag{8}$$

Since C_t is negative definite, there exists a $\delta > 0$ such that $w_t \cdot (C_t w_t) \leq -\delta \|w_t\|^2$. Also, since $\|C_t\| \|w_t\| \geq \|C_t w_t\|$ for any C_t, w_t:

$$\|w_{t+1}\|^2 = \|w_t\|^2 - 2\beta\delta \|w_t\|^2 + \beta^2 \|C_t\|^2 \|w_t\|^2 \tag{9}$$
$$\|w_{t+1}\|^2 = (1 - 2\beta\delta + \beta^2 \|C_t\|^2) \|w_t\|^2 \tag{10}$$

If $\beta < \frac{2\delta}{\|C_t\|^2}$, then selecting $\zeta = 2\beta\delta - \beta^2 \|C_t\|^2$ establishes the theorem. ■

The overall proof harnesses this rapid progress towards $-(C_t)^{-1} d_t$, and together with the fact that C_t and d_t converge, this yields the desired result.

Theorem 2 *(Theorem 2 in the paper) Given Assumption 1, y_t converges to $-(C^*)^{-1} d^*$ with probability 1. Formally, for all $\Delta > 0$ there exists a T such that with probability at least $1 - \Delta$, for all $t > T$:*

$$\left\|y_t + (C^*)^{-1} d^*\right\| \leq \Delta \tag{11}$$

Proof: In order to simplify the proof, we assume $d^* = 0$. A similar argument can be made when this is not the case. Observe that, for every $\epsilon > 0$, there exists a T such that, with probability at least $1 - \epsilon$ the following hold for all $t > T$ (which we will call event E):

1. $\|C_t - C^*\| < \epsilon$,
2. $\|d_t\| = \|d_t - d^*\| < \epsilon$, and
3. $\beta_t < \epsilon$.

In the following, we are assuming event E occurs. Thus, we can basically wait for C_t and d_t to get arbitrarily close to C^* and d^* (and stay there). This also allows for us to define a $\beta > 0$, and a sequence $k_1, k_2, \ldots$ (where $k_1 > T$) such that $\sum_{t=k_m}^{k_{m+1}-1} \beta_t = \bar{\beta}_m$, where $\beta \leq \bar{\beta}_m \leq 2\beta$.

Following the technique of [1], we define $q_m = y_{k_m}$. We then break down the transition from q_m to q_{m+1} into various quantities:

$$q_{m+1} = q_m + \sum_{t=k_m}^{k_{m+1}-1} \beta_t R_t (C_t y_t + d_t) \tag{12}$$

$$\begin{aligned} q_{m+1} = q_m &+ \sum_{t=k_m}^{k_{m+1}-1} \beta_t C^* q_m \\ &+ \sum_{t=k_m}^{k_{m+1}-1} \beta_t d_t \\ &+ \sum_{t=k_m}^{k_{m+1}-1} \beta_t (R_t - I) C^* (q_m) \\ &+ \sum_{t=k_m}^{k_{m+1}-1} \beta_t R_t (C_t - C^*) q_m \\ &+ \sum_{t=k_m}^{k_{m+1}-1} \beta_t R_t C_t (y_t - q_m) \end{aligned} \tag{13}$$

We can write this in the form $q_{m+1} = q_m + g_{1,m} + g_{2,m} + g_{3,m} + g_{4,m} + g_{5,m}$ where $g_{1,m} \dots g_{5,m}$ are the five sums on the right hand side of Equation 13.

We first prove a bound on $\|q_m + g_{1,m}\|$:

$$\|q_m + g_{1,m}\|^2 = \left\| q_m + \sum_{t=k_m}^{k_{m+1}-1} \beta_t C^* q_m \right\|^2 \tag{14}$$

$$= \left\| q_m + \bar{\beta}_m C^* q_m \right\|^2 \tag{15}$$

$$= q_m^2 + 2 q_m \cdot (\bar{\beta}_m C^* q_m) + (\bar{\beta}_m C^* q_m)^2 \tag{16}$$

Define $K = \max(M, \|C^*\| + \epsilon)$, such that K is a bound on all $\|R_t\|$, $\|C_t\|$, and $\|C^*\|$. Since C^* is negative definite, there exists a $\delta > 0$ such that for all x, $x \cdot (C^* x) \leq -\delta \|x\|^2$. Thus, the above can be reduced to:

$$\|q_m + g_{1,m}\|^2 \leq \|q_m\|^2 - 2\bar{\beta}_m \delta \|q_m\|^2 + (\bar{\beta}_m)^2 K^2 \|q_m\|^2 \tag{17}$$

$$\leq \|q_m\|^2 (1 - 2\bar{\beta}_m \delta + (\bar{\beta}_m)^2 K^2) \tag{18}$$

For a sufficiently small $\bar{\beta}_m$ (which can be guaranteed by a sufficiently small β), this is smaller than $\|q_m\|^2$. The remainder of the terms will do more harm than good.

$$\|g_{2,m}\| = \left\| \sum_{t=k_m}^{k_{m+1}-1} \beta_t R_t d_t \right\| \tag{19}$$

$$\leq \sum_{t=k_m}^{k_{m+1}-1} \beta_t \|R_t d_t\| \tag{20}$$

$$\leq \sum_{t=k_m}^{k_{m+1}-1} \beta_t K \epsilon \tag{21}$$

$$\leq \bar{\beta}_m K \epsilon \tag{22}$$

Observe that the bound in Equation 22 is not dependent on $\|q_m\|$. This is somewhat unfortunate, in that it means that we cannot *directly* prove $\|q_m\|$ approaches zero: instead, we prove that it becomes small, in terms related to ϵ.

$g_{3,m}$ has a more probabilistic flavor, in the sense that it is possible for it to be quite large.

$$\|g_{3,m}\| = \left\| \sum_{t=k_m}^{k_{m+1}} \beta_t (R_t - I) C^* q_m \right\| \tag{23}$$

$$\leq \sum_{t=k_m}^{k_{m+1}} \beta_t \|R_t - I\| \|C^*\| \|q_m\| \tag{24}$$

$$\leq (\bar{\beta}_m) K (K+1) \|q_m\| \tag{25}$$

In order to bound $g_{3,m}$ more tightly, we need to bound the size of $\left\| \sum_{t=k_m}^{k_{m+1}-1} \beta_t (R_t - I) \right\| < \epsilon_2$. This we can do probabilistically. We can achieve this by bounding each component of this matrix sum independently. First, observe that, for any i, j, $|(R_t - I)_{i,j}| \leq K + 1$. Define $x_{m,i,j} = \sum_{t=k_m}^{k_{m+1}-1} \beta_t (R_t - I)_{i,j}$. This allows us to use Azuma's Inequality:

$$\Pr[|x_{m,i,j}| \geq \epsilon_2] \leq 2 \exp(-2\epsilon_2^2 / (2\beta(K+1)^2 \beta_{k_m})) \tag{26}$$

Observe that if $\beta_t = vt^{-\mu}$, then:

$$\Pr[|x_{m,i,j}| \geq \epsilon_2] \leq 2 \exp(-2(vk_m^{\mu} \epsilon_2^2 / (2\beta(K+1)^2)) \tag{27}$$

We can observe that this is decreasing nearly exponentially with respect to m, so that by Lemma 5:

$$\Pr[\forall m, i, j : |x_{m,i,j}| \geq \epsilon_2] < \infty \tag{28}$$

For any $\epsilon_2 > 0$, we can choose an m^* such that:

$$\Pr[\forall m > m^*, \forall i, j : |x_{m,i,j}| < \epsilon_2] > 1 - \epsilon_2 \tag{29}$$

We will call this event E_2. Given E and E_2:

$$\|g_{3,m}\| \leq \left\| \sum_{t=k_m}^{k_{m+1}} \beta_t (R_t - I) C^* q_m \right\| \tag{30}$$

$$\leq \left\| \sum_{t=k_m}^{k_{m+1}} \beta_t (R_t - I) \right\| \|C^*\| \|q_m\| \tag{31}$$

$$\leq n\epsilon_2 K \|q_m\| \tag{32}$$

The next term is relatively simple:

$$\|g_{4,m}\| = \left\| \sum_{t=k_m}^{k_{m+1}-1} \beta_t R_t (C_t - C^*) q_m \right\| \tag{33}$$

$$\leq \sum_{t=k_m}^{k_{m+1}-1} \beta_t \|R_t (C_t - C^*) q_m\| \tag{34}$$

$$\leq \sum_{t=k_m}^{k_{m+1}-1} \beta_t K \epsilon \|q_m\| \tag{35}$$

$$\leq \bar{\beta}_m K \epsilon \|q_m\| \tag{36}$$

Finally, we have to bound the effect of the drift of the algorithm upon the direction of change.

Lemma 3 *For all t such that $k_m \leq t < k_{m+1}$:*

$$\|y_t - q_m\| \leq \bar{\beta}_m K^2 (\|q_m\| + \bar{\beta}_m K \epsilon) \exp(\bar{\beta}_m K^2) \tag{37}$$

Proof: Observe that, for any $x \in \mathbf{R}^n$:

$$\|x + \beta_t R_t (C_t x + d_t)\| \leq \|(I + \beta_t R_t C_t) x\| + \|\beta_t R_t d_t\| \tag{38}$$

$$\leq (1 + \beta_t K^2) \|x\| + \beta_t K \epsilon \tag{39}$$

Now, applying this operator recursively:

$$\|y_t\| \leq \|q_m\| \prod_{t'=k_m}^{t-1} (1 + \beta_{t'} K^2) \tag{40}$$

$$+ \sum_{t''=k_m}^{t-1} \beta_{t''} K\epsilon \prod_{t'=t''+1}^{t-1} (1 + \beta_{t'} K^2) \tag{41}$$

$$\leq (\|q_m\| + \bar{\beta}_m K\epsilon) \exp(\bar{\beta}_m K^2) \text{ if } t < k_{m+1} \tag{42}$$

We can now compute the difference between y_t and q_m (assuming $t < k_{m+1}$).

$$\|y_t - q_m\| \leq \sum_{t'=k_m}^{t-1} \beta_{t'} \|R_{t'} C_{t'}\| \|y_{t'}\| \tag{43}$$

$$\leq \sum_{t'=k_m}^{t-1} \beta_{t'} K^2 (\|q_m\| + \bar{\beta}_m K\epsilon) \exp(\bar{\beta}_m K^2) \tag{44}$$

$$\leq \bar{\beta}_m K^2 (\|q_m\| + \bar{\beta}_m K\epsilon) \exp(\bar{\beta}_m K^2) \tag{45}$$

■

We now continue with the proof of Theorem 2. The last term $g_{5,m}$ can be bounded by:

$$\|g_{5,m}\| \leq \sum_{t=k_m}^{k_{m+1}-1} \beta_t \|R_t C_t\| \|y_t - q_m\| \tag{46}$$

$$\leq \sum_{t=k_m}^{k_{m+1}-1} \beta_t K^2 \bar{\beta}_m K^2 (\|q_m\| + \bar{\beta}_m K\epsilon) \exp(\bar{\beta}_m K^2) \tag{47}$$

$$\leq \bar{\beta}_m K^2 \bar{\beta}_m K^2 (\|q_m\| + \bar{\beta}_m K\epsilon) \exp(\bar{\beta}_m K^2) \tag{48}$$

$$\leq (\bar{\beta}_m)^2 K^4 (\|q_m\| + \bar{\beta}_m K\epsilon) \exp(\bar{\beta}_m K^2) \tag{49}$$

Observe that we have only assumed that $\beta > 0$. By making the assumption that $\exp(2\beta K^2) \leq 2$, $2K^2\beta \leq \delta$ (such that $-2\bar{\beta}_m\delta + (\bar{\beta}_m)^2 K^2 \leq -\bar{\beta}_m\delta$), and remembering that $\beta \leq \bar{\beta}_m \leq 2\beta$, we can summarize the above results (assuming

E):

$$\|q_m + g_{1,m}\|^2 \leq \|q_m\|^2 (1 - \beta\delta) \tag{50}$$
$$\|g_{2,m}\| \leq 2\beta K\epsilon \tag{51}$$
$$\|g_{3,m}\| \leq n\epsilon_2 K \|q_m\| \tag{52}$$
$$\|g_{3,m}\| \leq 2\beta K(K+1) \|q_m\| \tag{53}$$
$$\|g_{4,m}\| \leq 2\beta K\epsilon \|q_m\| \tag{54}$$
$$\|g_{5,m}\| \leq 8\beta^2 K^4 (\|q_m\| + 2\beta K\epsilon) \tag{55}$$

In order to make these results a little cleaner, we introduce a variable L:

$$L = \max(2K\sqrt{\epsilon}, 2K^2\sqrt{\epsilon}, 2\sqrt{\beta}K(K+1), 2K\epsilon, 8\beta K^4, 16\beta^2 K^5\sqrt{\epsilon}). \tag{56}$$

Observe that $\lim_{\beta,\epsilon\to 0} L = 0$, so that we can choose L to be arbitrarily small (with respect to δ and M). This allows us to state the above results in a remarkably cleaner way:

$$\|q_m + g_{1,m}\|^2 \leq \|q_m\|^2 (1 - \beta\delta) \tag{57}$$
$$\|g_{2,m}\| \leq \beta L\sqrt{\epsilon} \tag{58}$$
$$\|g_{3,m}\| \leq n\epsilon_2 K \|q_m\| \tag{59}$$
$$\|g_{3,m}\| \leq \sqrt{\beta} L \|q_m\| \tag{60}$$
$$\|g_{4,m}\| \leq \beta L \|q_m\| \tag{61}$$
$$\|g_{5,m}\| \leq \beta L \|q_m\| + \beta L\sqrt{\epsilon} \tag{62}$$

Note that the bound on $g_{3,m}$ is weaker than the others, because as $\beta \to 0$, it is the case that $\sqrt{\beta} > \beta$. Thus, it will require special attention. Now, we bound q_{m+1}:

$$\|q_{m+1}\|^2 \leq \|q_m + g_{1,m}\|^2 + \|g_{2,m} + g_{4,m} + g_{5,m}\|^2 \tag{63}$$
$$+ 2\|g_{2,m} + g_{4,m} + g_{5,m}\| \|q_m + g_{1,m}\| + \|g_{3,m}\|^2 \tag{64}$$
$$+ 2\|g_{2,m} + g_{4,m} + g_{5,m}\| \|g_{3,m}\| + 2g_{3,m} \cdot (q_m + g_{1,m}) \tag{65}$$

We observe that $\|q_m + g_{1,m}\| \leq \|q_m\|$ and $\|g_{2,m} + g_{4,m} + g_{5,m}\| \leq 2\beta L(\|q_m\| + \sqrt{\epsilon})$.

$$\begin{aligned}\|q_{m+1}\|^2 \leq\ & \|q_m\|^2 (1 - \beta\delta) + (2\beta L(\|q_m\| + \sqrt{\epsilon}))^2 \\ & + 2(2\beta L(\|q_m\| + \sqrt{\epsilon})) \|q_m\| + \|g_{3,m}\|^2 \\ & + 2(2\beta L(\|q_m\| + \sqrt{\epsilon})) \|g_{3,m}\| + 2g_{3,m} \cdot (q_m + g_{1,m})\end{aligned} \tag{66}$$

$$\begin{aligned}\|q_{m+1}\|^2 \leq\ & \|q_m\|^2 (1 - \beta\delta) + 4\beta^2 L^2 (\|q_m\| + \sqrt{\epsilon})^2 \\ & + 4\beta L(\|q_m\| + \sqrt{\epsilon}) \|q_m\| + \|g_{3,m}\|^2 \\ & + 4\beta L(\|q_m\| + \sqrt{\epsilon}) \|g_{3,m}\| + 2g_{3,m} \cdot (q_m + g_{1,m})\end{aligned} \tag{67}$$

Now, we can use the weak bound on $g_{3,m}$ on two of its occurrences.

$$\begin{aligned}\|q_{m+1}\|^2 \leq \|q_m\|^2(1-\beta\delta) + 4\beta^2L^2(\|q_m\| + \sqrt{\epsilon})^2 \\ + 4\beta L(\|q_m\| + \sqrt{\epsilon})\|q_m\| + \beta L^2 \\ + 4\beta L(\|q_m\| + \sqrt{\epsilon})\sqrt{\beta}L + 2g_{3,m}\cdot(q_m + g_{1,m})\end{aligned} \tag{68}$$

$$\begin{aligned}\|q_{m+1}\|^2 \leq \|q_m\|^2(1-\beta\delta) + 4\beta^2L^2(\|q_m\| + \sqrt{\epsilon})^2 \\ + 4\beta L(\|q_m\| + \sqrt{\epsilon})\|q_m\| + \beta L^2\|q_m\|^2 \\ + 4\beta^{3/2}L^2(\|q_m\| + \sqrt{\epsilon})\|q_m\|\end{aligned} \tag{69}$$

$$\begin{aligned}\|q_{m+1}\|^2 \leq \max(\|q_m\|, \sqrt{\epsilon})^2[1-\beta\delta + 16\beta^2L^2 + 8\beta L + \beta L^2 + 8\beta^{3/2}L^2] \\ + 2g_{3,m}\cdot(q_m + g_{1,m})\end{aligned} \tag{70}$$

If $16\beta L^2 + 8L + L^2 + 8\sqrt{\beta}L^2 \leq \delta/2$, then:

$$\|q_{m+1}\|^2 \leq \max(\|q_m\|, \sqrt{\epsilon})^2[1-\beta\delta/2] + 2g_{3,m}\cdot(q_m + g_{1,m}) \tag{71}$$

We now fix β and ϵ and approach the last term. If E_2 holds and $m > m^*$, then:

$$\|q_{m+1}\|^2 \leq \max(\|q_m\|, \sqrt{\epsilon})^2[1-\beta\delta/2] + 2nK\epsilon_2\|q_m\|^2 \tag{72}$$

Assuming that $\epsilon_2 < \frac{\beta\delta/4}{2nK}$, then:

$$\|q_{m+1}\|^2 \leq \max(\|q_m\|, \sqrt{\epsilon})^2[1-\beta\delta/4] \tag{73}$$

Thus, if E and E_2 hold, for all $m > m^*$, $\|q_m\|$ will contract until it reaches ϵ. Therefore, if $\epsilon + \epsilon_2 < \Delta$, the result will hold. ■

Of course, we have to map iLSTD to such a process:

1. $y_t = \theta_t$,
2. $\beta_t = t\alpha/n$,
3. $C_t = -\mathbf{A}_t/t$,
4. $d_t = \mathbf{b}_t/t$, and
5. R_t is a matrix, where there is an n on the diagonal in position (k,k) (where k is the randomly selected dimension on round t) and zeroes everywhere else.

Theorem 4 *(Theorem 3 in the paper) If the Markov decision process is finite, iLSTD(λ) with a uniform random feature selection mechanism converges to the same result as TD(λ).*

Proof: Observe that for any $\epsilon > 0$ there exists a Δt such that $\Delta \mathbf{A}$ can be approximated to within ϵ by by only considering the last Δt steps. Since the distribution over the last Δt time steps at time $T+1$ can be determined precisely by the last Δt time steps at time T, and there are a finite number of such states, the average of the $\Delta \mathbf{A}$ (i.e., $-C_t$) must converge, at least within ϵ. Since this holds for any ϵ, C_t converges. A similar argument holds for d_t.

Since α is decreasing like $t^{-(1+\mu)}$ where $\mu \in (0,1]$, β satisfies the above properties. ∎

References

[1] Dmitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Lemma 5 *For any $K, \alpha > 0$,*

$$\sum_{t=1}^{\infty} \exp(-Kt^{-\alpha}) < \infty \tag{74}$$

Proof: Observe that $\sum_{t=1}^{\infty} t^{-2} < \infty$. Moreover, $t^{-2} = \exp(-2 \ln t)$. Observe that for sufficiently large t, $-2 \ln t > -Kt^{-\alpha}$. Therefore, the tail of the sum has a finite value, making the sum finite. ∎