Temporal Difference Learning by Direct Preconditioning

Hengshuai Yao
Department of Computing Science, University of Alberta, Edmonton, AB Canada T6G2E8
HENGSHUA@CS.UALBERTA.CA

Shalabh Bhatnagar
Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India 560 012
SHALABH@CSA.IISC.ERNET.IN

Csaba Szepesvári
Department of Computing Science, University of Alberta, Edmonton, AB Canada T6G2E8
SZEPESVA@CS.UALBERTA.CA

Abstract
We propose a new class of algorithms that directly precondition the TD update. We then focus on a new preconditioned algorithm and prove its convergence. Empirical results on the new algorithm shall be presented in a detailed version of this paper.

1. Direct Preconditioned TD algorithms

Previous work (Yao & Liu, 2008) relates LSTD, LSPE, and iLSTD via a class of Preconditioned TD (PTD) algorithms. This paper explores yet another class of preconditioned algorithms.

We consider on-policy policy evaluation using a linear function approximation (Sutton & Barto, 1998). For each state \( i \), there is a corresponding feature vector \( \phi(i) \in \mathbb{R}^n \) where \( n < N \). On a transition from state \( s_t \) to state \( s_{t+1} \), we obtain a reward \( r_t \), and apply TD(0):

\[
\theta_{t+1} = \theta_t + \alpha_t \phi_t \delta_t,
\]

where \( \phi_t = \phi(s_t) \), \( \delta_t = r_t + \gamma \theta_t^T \phi_{t+1} - \theta_t^T \phi_t \) and \( \alpha_t \) is a positive scalar. The term, \( \delta_t \), is usually referred as the TD-update. For the ergodic problem, TD(0) converges to a solution of the system of equations

\[
E[\delta \phi] = A \theta^* + b = 0, \quad A = E[\phi_t (\phi_{t+1} - \phi_t)^T], \quad b = E[\phi_t r_t].
\]

Note that the PTD algorithms in (Yao & Liu, 2008) take the following form:

\[
\theta_{t+1} = \theta_t + \alpha_t P_t^{-1} (A_t \theta_t + b_t),
\]

where \( P_t \) is an invertible preconditioner matrix, and \( A_t, b_t \) are some estimations of \( A, b \) respectively. Here we propose another class of preconditioned TD algorithms, cast as

\[
\theta_{t+1} = \theta_t + \alpha_t P_t^{-1} \delta_t \phi_t. \tag{1}
\]

The new class of algorithms precondition the TD-update directly, rather than the residual vector, \( A_t \theta_t + b_t \). In (1), if we use \( P_t = I \), we recover TD; however, if \( P_t = D_t \), where \( D_t \) is some estimation of \( D = E[\phi_t^T \phi_t] \), we obtain the Fixed-point Kalman Filter (FPKF) (Choi & Van Roy, 2006); and if \( P_t = -A_t \), we get an algorithm that is reminiscent of Newton method, which we call the Newton TD (NTD) algorithm.

2. The Newton TD Algorithm

The algorithms updates according to

\[
\theta_{t+1} = \theta_t - \alpha_t A_t^{-1} \delta_t \phi_t. \tag{2}
\]
where $A_t^{-1}$ are recursively obtained as

$$A_{t+1}^{-1} = \frac{1}{1-\beta_t} \left( A_t^{-1} - \frac{\beta_t A_t^{-1} \phi_t (\gamma \phi_{t+1} - \phi_t) A_t^{-1} \phi_t}{1-\beta_t + \beta_t (\gamma \phi_{t+1} - \phi_t) A_t^{-1} \phi_t} \right)$$

We will make the following two assumptions:

(A1) The step-sizes $\alpha_t$, $\beta_t$, $t \geq 0$ satisfy $a(t), b(t) > 0$ for all $t$. Further, $\alpha_t \sim t$ as $t \to \infty$, $\sum_t \alpha_t = \beta_t = \infty$, $\sum_t \beta_t^2 < \infty$, $\alpha_t = O(\beta_t)$.

(A2) The iterates $A_t$, $t \geq 1$ satisfy $\sup_t \| A_t \|$, $\sup_t \| A_t^{-1} \| < \infty$.

(A1) essentially implies that we have decreasing step-size sequences and in addition $\alpha_t \to 0$ faster than $\beta_t$ does. In effect, it implies that the recursion governed by $\beta_t$ is faster as opposed to the one governed by $\alpha_t$. (A2) ensures that the iterates $A_t$, $A_t^{-1}$, $t \geq 1$ do not blow up as $t \to \infty$. A sufficient condition for (A2) is the following: Let there exist scalars $c_1, c_2 > 0$ with $c_1 < c_2$ such that $c_1 \parallel z \parallel^2 \leq |Re(z^T A_t z)| \leq c_2 \parallel z \parallel^2$, for all $t \geq 0$, $z \in \mathbb{R}^n$. The above implies that the real parts of the eigenvalues of $A_t$ remain either in the interval $[c_2, c_1]$ or else in the interval $[c_1, c_2]$. Thus the real parts of the eigenvalues of $A_t^{-1}$ remain either in the interval $[-\frac{1}{c_2}, -\frac{1}{c_1}]$ or else in the interval $[\frac{1}{c_2}, \frac{1}{c_1}]$. This will ensure that the eigenvalues of $A_t^{-1}$ remain absolutely uniformly bounded both from above as well as away from zero.

For any $n \times n$ matrix $B$, we define its norm $\| B \|$ as the norm induced from the corresponding vector norm and is defined as $\| B \| = \max_{\parallel x \parallel = 1} \| Bx \|$. We have the following convergence result.

**Theorem 1** (Convergence of NTD). Under assumptions (A1)-(A2), $\theta_t \to \theta^*$ as $t \to \infty$ with probability one, where $\theta^* = -A^{-1}b$.

**Proof.** The proof relies on a two-timescale analysis (see (A1)). Note that the recursion (3) corresponds to the faster recursion while (2) is the slower one. Thus from the timescale of (2), i.e., that corresponding to $\{\alpha_t\}$, recursion (3) appears equilibrated while from the other timescale corresponding to $\{\beta_t\}$, the recursion (2) is quasi-static. Consider now (3). Using a standard convergence analysis under (A2), it can be seen that $A_t \to A$ as $t \to \infty$. Now note that $\| A_t^{-1} - A^{-1} \| \leq \| A^{-1} (A_t - A) A_t^{-1} \| \leq \| A^{-1} \| \sup_t \| A_t^{-1} \| \| A_t - A \| \to 0$ as $t \to \infty$, in lieu of (A2) and the above. On the other hand, since $\alpha_t = o(\beta_t)$, one can write (2) as $\theta_{t+1} = \theta_t - \beta_t \xi_t$, where $\xi_t = \left( \frac{\alpha_t}{\beta_t} A_t^{-1} \beta_t \phi_t \right) = o(1)$ by (A1). Hence, along the faster timescale (i.e., the one corresponding to $\{\beta_t\}$), $A_t^{-1} \to A^{-1}$, while $\theta_t \to \theta$ (i.e., the latter is quasi-static). Next consider recursion (2) along its timescale (i.e., the slower one corresponding to $\{\alpha_t\}$) with $A_t^{-1}$ equilibrated. Thus consider $\theta_{t+1} = \theta_t - \alpha_t A^{-1} \delta_t \phi_t$. Let $F_t = \sigma(\phi_x, s \prec t, t \geq 1)$. Now rewrite the above as $\theta_{t+1} = \theta_t - \alpha_t A^{-1} E[\delta_t \phi_t | F_t] - \alpha_t A^{-1} (E[\delta_t \phi_t | F_t])$. Define the sequence $\{N_t\}$ as follows: $N_t = \sum_{s=0}^{t} \alpha_s A^{-1} (\delta_s \phi_s - E[\delta_s \phi_s | F_s])$. It is easy to see that $\{N_t, F_t\}$ is a martingale sequence. By the martingale convergence theorem, under (A1)-(A2) and the fact that $\phi_s$ are uniformly bounded features, one can see that $\{N_t, F_t\}$ is also convergent. Thus, for any $T > 0$ with $n_T = \min\{m \geq n | \sum_{r=m}^{n} \alpha_r \geq T\}$, we have that $\sum_{n_T}^{n} \alpha_s A^{-1} (\delta_s \phi_s - E[\delta_s \phi_s | F_s]) \to 0$ a.s. as $n \to \infty$. Consider now the ordinary differential equation (ODE)

$$\dot{\theta} = -A^{-1} (A\theta + b) = -(\theta + A^{-1}b).$$

Let $h(\theta) = -(\theta + A^{-1}b)$ i.e., the RHS of (4). Then $h(\cdot)$ is a Lipschitz continuous function implying that the ODE (4) is well posed. Further, $\theta^* = -A^{-1}b$ is the unique asymptotically stable equilibrium for (4). Now let $h(\infty) = \lim_{r \to \infty} h(r\theta)/r = -\theta$. Consider an associated ODE $\tilde{\theta} = h(\infty)(\theta) = -\theta$. For the latter ODE, the origin is an asymptotically stable equilibrium. The recursion (2) is now uniformly bounded from Theorem 2.1 of (Borkar & Meyn, 2000). The claim now follows as a consequence of the Hirsch’s lemma (cf. Theorem 1, pp.339 of (Hirsch, 1989)) in a similar manner as Theorem 2.2 of (Borkar & Meyn, 2000). This completes the proof.

**References**


