Convergence of Value-Based Reinforcement Learning:

Advances and Open Problems

Shangtong Zhang

Assistant Professor
Department of Computer Science
University of Virginia

Value-based RL can be viewed as the combination of stochastic approximation and dynamic programming

- Goal: $\mathcal{T}(w_*) = w_*$
- Incremental learning: $w_{t+1} = w_t + \alpha_t(\mathcal{T}(w_t) - w_t)$
- Stochastic approximation: $w_{t+1} = w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t)$
  $\mathbb{E}[T(w, y)] = \mathcal{T}(w)$
- Different types of noise $\{Y_t\}$:
  - i.i.d.
  - Markov chain
  - time-inhomogeneous Markov chain

Following Occam's razor, we analyze the algorithms with minimal assumptions / modifications

- Don't add projection $w_{t+1} = \prod(w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t))$
- Don't use "local clock" $\alpha_t \to \alpha_{\nu(Y_{t+1},t)}$
- Don't assume linear MDP
- Don't add regularizer (e.g., ridge)
- ...

Case study: linear Q-learning

$$w_{t+1} = w_t + \alpha_t(R_{t+1} + \gamma \max_{a'} x(S_{t+1}, a')^\top w_t - x(S_t, A_t)w_t)x(S_t, A_t)$$

$$A_t \sim \mu_{w_t}(\cdot|S_t)$$

SA Formulation

$$Y_t = (S_t, A_t, S_{t+1}) \quad \text{time-inhomogeneous but finite}$$

$$T(w, (s, a, s')) - w = (r(s, a) + \gamma \max_{a'} x(s', a')^\top w - x(s, a)^\top w_t)x(s, a)$$

Linear Q-learning does NOT diverge

when an ϵ-softmax behavior policy with an adaptive temperature is used.

$$\mu_w(a|s) = \frac{\epsilon}{|\mathcal{A}|} + (1-\epsilon)\frac{\exp\left(\tau_w x(s,a)^\top w\right)}{\sum_b \exp(\tau_w x(s,b)^\top w)}$$

$$\tau_w = \begin{cases} \frac{\tau_0}{\|w\|_2} & \|w\|_2 > 1 \\ \tau_0 & \text{otherwise} \end{cases}$$

- Meyn (2024): $\limsup_t \|w_t\| < C$ a.s.
- Liu et al. (2025b): $\mathbb{E}\left[\|w_t\|^2\right] \leq C + g(t)$

Case study: linear TD(λ)

$$
\begin{aligned}
e_t &= \lambda\gamma e_{t-1} + x(S_t) \\
w_{t+1} &= w_t + \alpha_t(R_{t+1} + \gamma x(S_{t+1})^\top w_t - x(S_t)^\top w_t)e_t
\end{aligned}
$$

SA Formulation

$$
\begin{aligned}
Y_t &= (S_t, A_t, S_{t+1}, e_t) \quad \text{infinite but compact} \\
T(w, (s, a, s', e)) - w &= (r(s, a) + \gamma x(s')^\top w - x(s)^\top w)e
\end{aligned}
$$

Does linear TD(λ) converge?

- With linearly independent features, yes (Tsitsiklis and Roy, 1996)
- But without linearly independent features? hmm...
  - asked by Peter (1992); Tsitsiklis and Roy (1996, 1999)
  - $\lim_{t\to\infty} x(s)^\top w_t = \hat{v}(s)$ a.s. (Wang and Zhang, 2024)
  - $\lim_{t\to\infty} w_t(\omega) = w_*(\omega)$ (still open)

Case study: GTD(λ)

$$
\begin{aligned}
e_t &= \lambda\gamma \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} e_{t-1} + x(S_t) \\
w_{t+1} &= w_t + \alpha_t \dots
\end{aligned}
$$

SA Formulation

$$Y_t = (S_t, A_t, S_{t+1}, e_t) \quad \text{infinite}$$

$\{Y_t\}$ behaves really poor (Yu, 2012, 2017)

- unbounded second moments
- unbounded almost surely

Case study: average reward TD

$$
\begin{aligned}
J_{t+1} &= J_t + \beta_t(R_{t+1} - J_t) \\
v_{t+1}(S_t) &= v_t(S_t) + \alpha_t(R_{t+1} - J_t + v_t(S_{t+1}) - v_t(S_t))
\end{aligned}
$$

Average reward TD converges to a sample path dependent fixed point

- Tsitsiklis and Roy (1999): $\lim_{t\to\infty} d(v_t, V_*) = 0$ a.s.
- Blaser and Zhang (2024): $\lim_{t\to\infty} v_t(\omega) = v_*(\omega)$

ω is a sample path, i.e., a realization of $S_0, A_0, S_1, R_1, A_1, \ldots$.

Case study: differential Q learning (Wan et al., 2021)

$$
\begin{aligned}
\delta_t &= R_{t+1} - J_t + \max_{a'} q_t(S_{t+1}, a') - q_t(S_t, A_t) \\
J_{t+1} &= J_t + \alpha_t \eta \delta_t \\
q_{t+1}(S_t, A_t) &= q_t(S_t, A_t) + \alpha_t \delta_t
\end{aligned}
$$

Does differential Q learning converge?

- with local clock ($\alpha_t \to \alpha_{\nu(Y_{t+1},t)}$), yes
- without local clock, hmm... unknown, rank 1 pertubation

Three tools are commonly used

$$w_{t+1} = w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t)$$

- ODE-based analysis (Benveniste et al., 1990; Kushner and Yin, 2003; Borkar, 2009)
  - Liu et al. (2025a)
- "almost" supermartingales (Robbins and Siegmund, 1971)
  - Qian et al. (2024)
- Krasnoselskii-Mann (KM) iterations (Krasnosel'skii, 1955; Cominetti et al., 2014; Bravo and Cominetti, 2022)
  - Blaser and Zhang (2024)

ODE approaches connect stochastic and discrete iterates with deterministic and continuous trajectories

$$
\begin{aligned}
w_{t+1} &= w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t) \\
\frac{\mathrm{d}w(t)}{\mathrm{d}t} &= \mathcal{T}(w(t)) - w(t)
\end{aligned}
$$

Previous ODE approaches have various restrictions

- Kushner and Yin (2003) require stability as a priori
  - Assume $\sup_t \|w_t\| < \infty$ a.s.
  - Add projection $w_{t+1} = \prod (w_t + \alpha_t (T(w_t, Y_{t+1}) - w_t))$
- Borkar (2009) requires i.i.d. noise or local clock
  - $\{Y_t\}$ i.i.d.
  - $\{Y_t\}$ Markovian, $\alpha_t \to \alpha_{\nu(Y_{t+1}, t)}$
- Benveniste et al. (1990) require Poisson's equation; Borkar et al. (2021) require Lyapunov drift condition V4
  - Finite $\{Y_t\}$ is usually fine
  - But off-policy traces in GTD(λ) (Yu, 2017) or ETD(λ) (Yu, 2015) do not work. $Y_t = (S_t, A_t, S_{t+1}, e_t)$

Liu et al. (2025a) require only Law of Large Numbers!

Law of large numbers on Markov chains

$$\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} f(Y_t) = \mathbb{E}[f(y)] \quad \text{a.s. (LLN)}$$

The $\{Y_t\}$ in both GTD(λ) and ETD(λ) satisfy this LLN (Yu, 2012, 2015, 2017).

$$w_{t+1} = w_t + \alpha_t (T(w_t, Y_{t+1}) - w_t)$$

$$w_t \to w_*$$

LLN is accompanied by a bunch of other results characterizing the fluctuation of the empirical mean

Suppose $\{\zeta_t\}$ are i.i.d.

- LLN: $\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} f(\zeta_t) = \mathbb{E}[f(\zeta)]$ a.s.
- Central Limit Theorem (CLT)
- Functional CLT (FCLT)
- Almost sure convergence rate (e.g., LIL)
- High Probability Concentration (e.g., Hoeffding's inequality)
- L^p convergence rate

Can we get those characterization for SA with Markov noise?

$$w_{t+1} = w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t)$$

- LLN: the ODE approach (Borkar et al., 2021; Liu et al., 2025a)
- CLT: Borkar et al. (2021)
- FCLT: Borkar et al. (2021)

Can we get those characterization for SA with Markov noise?

$$w_{t+1} = w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t)$$

Qian et al. (2024) get almost sure convergence rates, maximal concentration with exponential tails, and L^p convergence rates simultaneously

- Almost sure convergence rates:

$$\lim_{t\to\infty} \frac{\|w_t - w_*\|}{g(t)} = 0 \quad \text{a.s.}$$

- Concentration:

$$\Pr\Big(\|w_t - w_*\|^2 \leq Cg(t)\log(1/\delta)\ \forall t\Big) \geq 1 - \delta$$

- L^p convergence rates:

$$\mathbb{E}\big[\|w_t - w_*\|^p\big] = g(t)$$

Qian et al. (2024) are based on "almost" supermartingales and a skeleton iterates technique

- Supermartingale: $\mathbb{E}[M_{t+1}|M_1, \ldots, M_t] \leq M_t$
- Almost supermartingale (Robbins and Siegmund, 1971; Chen et al., 2023; Liu and Yuan, 2024):

$$\mathbb{E}[M_{t+1}|M_1, \ldots, M_t] \leq (1 - \alpha_t)M_t + \alpha_t^2$$

- Skeleton iterates – the iterates in a different timescale

$$\begin{aligned} w_{t+1} &= w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t) \\ w_{t_{m+1}} &= w_{t_m} + \alpha_{t_m}(\mathcal{T}(w_{t_m}) - w_{t_m} + \xi_m) \end{aligned}$$

KM iterations use a fox-and-hare race model to solve a recursion with nonexpansive operators.

- Cominetti et al. (2014)

$$w_{t+1} = w_t + \alpha_t(\mathcal{T}(w_t) - w_t)$$

$$w_m - w_n = \sum_{j=0}^{m} \sum_{k=m+1}^{n} \pi_j^m \pi_k^n (\mathcal{T} x_{j-1} - \mathcal{T} x_{k-1})$$

- Bravo and Cominetti (2022): i.i.d. noise

$$w_{t+1} = w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t)$$

- Blaser and Zhang (2025): Markovian noise

$$w_{t+1} = w_t + \alpha_t(T(w_t, Y_{t+1}) - w_t)$$

Acknowledgements

- Shuze Liu (UVA)
- Xinyu Liu (UVA)
- Zixuan Xie (UVA)
- Ethan Blaser (UVA)
- Jiuqi Wang (UVA)
- Xiaochi Qian (Oxford)
- Shuhang Chen (General Robotics)

References i

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer.

Blaser, E. and Zhang, S. (2024). Asymptotic and finite sample analysis of nonexpansive stochastic approximations with markovian noise. *ArXiv Preprint*.

Blaser, E. and Zhang, S. (2025). Asymptotic and finite sample analysis of nonexpansive stochastic approximations with markovian noise. *ArXiv Preprint*.

Borkar, V. (2009). *Stochastic approximation: a dynamical systems viewpoint*. Springer.

Borkar, V., Chen, S., Devraj, A., Kontoyiannis, I., and Meyn, S. (2021). The ode method for asymptotic statistics in stochastic approximation and reinforcement learning. *ArXiv Preprint*.

References ii

Bravo, M. and Cominetti, R. (2022). Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *ArXiv Preprint*.

Chen, Z., Maguluri, S. T., and Zubeldia, M. (2023). Concentration of contractive stochastic approximation: Additive and multiplicative noise. *ArXiv Preprint*.

Cominetti, R., Soto, J. A., and Vaisman, J. (2014). On the rate of convergence of krasnosel'skii-mann iterations and their connection with sums of bernoullis. *Israel Journal of Mathematics*.

Krasnosel'skii, M. A. (1955). Two remarks on the method of successive approximations. *Uspekhi matematicheskikh nauk*.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media.

References iii

Liu, J. and Yuan, Y. (2024). Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods. *Journal of Machine Learning Research*.

Liu, S., Chen, S., and Zhang, S. (2025a). The ODE method for stochastic approximation and reinforcement learning with markovian noise. *Journal of Machine Learning Research*.

Liu, X., Xie, Z., and Zhang, S. (2025b). Linear Q-learning does not diverge in L^2: Convergence rates to a bounded set.

Meyn, S. (2024). The projected bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*.

Peter, D. (1992). The convergence of td(λ) for general λ. *Machine Learning*.

References iv

Qian, X., Xie, Z., Liu, X., and Zhang, S. (2024). Almost sure convergence rates and concentration of stochastic approximation and reinforcement learning with markovian noise. *ArXiv Preprint*.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*.

Tsitsiklis, J. N. and Roy, B. V. (1996). Analysis of temporal-diffference learning with function approximation. In *IEEE Transactions on Automatic Control*.

Tsitsiklis, J. N. and Roy, B. V. (1999). Average cost temporal-difference learning. *Automatica*.

Wan, Y., Naik, A., and Sutton, R. S. (2021). Learning and planning in average-reward markov decision processes. In *Proceedings of the International Conference on Machine Learning*.

References v

Wang, J. and Zhang, S. (2024). Almost sure convergence of linear temporal difference learning with arbitrary features. *ArXiv Preprint*.

Yu, H. (2012). Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*.

Yu, H. (2015). On convergence of emphatic temporal-difference learning. In *Proceedings of the Conference on Learning Theory*.

Yu, H. (2017). On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *ArXiv Preprint*.