An Emphatic Approach to the Problem of Off-policy TD Learning

Rich Sutton

Rupam Mahmood

Martha White

Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science
University of Alberta, Canada

Temporal-Difference Learning with Linear Function Approximation

states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ policy $\pi(a|s) \doteq \mathbb{P}\{A_t = a | S_t = s\}$

transition prob matrix $[\mathbf{P}_\pi]_{ij} \doteq \sum_a \pi(a|i) p(j|i,a)$ where $p(j|i,a) \doteq \mathbb{P}\{S_{t+1} = j | S_t = i, A_t = a\}$

ergodic stationary distribution $[\mathbf{d}_\pi]_s \doteq d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t = s\} > 0$ $\quad \mathbf{P}_\pi^\top \mathbf{d}_\pi = \mathbf{d}_\pi$

return $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \quad 0 \le \gamma < 1$

value function $v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] \approx \mathbf{w}_t^\top \mathbf{x}(s)$

feature vectors $\mathbf{x}(s) \in \mathbb{R}^n \quad \forall s \in \mathcal{S}$

weight vector $\mathbf{w}_t \in \mathbb{R}^n \quad n \ll |\mathcal{S}|$

linear TD(0):

$$
\begin{aligned}
\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}(S_{t+1}) - \mathbf{w}_t^\top \mathbf{x}(S_t) \right) \mathbf{x}(S_t) \\
&= \mathbf{w}_t + \alpha \Big(\underbrace{R_{t+1}\mathbf{x}(S_t)}_{\mathbf{b}_t \in \mathbb{R}^n} - \underbrace{\mathbf{x}(S_t)\left(\mathbf{x}(S_t) - \gamma \mathbf{x}(S_{t+1})\right)^\top}_{\mathbf{A}_t \in \mathbb{R}^{n\times n}} \mathbf{w}_t \Big) \\
&= \mathbf{w}_t + \alpha(\mathbf{b}_t - \mathbf{A}_t \mathbf{w}_t) \\
&= (\mathbf{I} - \alpha \mathbf{A}_t)\mathbf{w}_t + \alpha \mathbf{b}_t.
\end{aligned}
$$

deterministic 'expected' update: $\bar{\mathbf{w}}_{t+1} \doteq (\mathbf{I} - \alpha \mathbf{A})\bar{\mathbf{w}}_t + \alpha \mathbf{b}$

$$
\begin{aligned}
\mathbf{A} \doteq \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] &= \lim_{t\to\infty} \mathbb{E}_\pi \left[\mathbf{x}(S_t)\left(\mathbf{x}(S_t) - \gamma \mathbf{x}(S_{t+1})\right)^\top \right] \\
&= \sum_s d_\pi(s)\, \mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \sum_{s'} [\mathbf{P}_\pi]_{ss'} \mathbf{x}(s') \right)^\top \\
&= \mathbf{X}^\top \underbrace{\mathbf{D}_\pi (\mathbf{I} - \gamma \mathbf{P}_\pi)}\mathbf{X},
\end{aligned}
$$

Stable if $\mathbf{A}$ is positive definite
i.e., if $\mathbf{y}^\top \mathbf{A} \mathbf{y} > 0, \ \forall \mathbf{y} \neq \mathbf{0}.$
Converges to $\lim_{t\to\infty} \bar{\mathbf{w}}_t = \mathbf{A}^{-1}\mathbf{b}.$

$$
\mathbf{X} \doteq \begin{bmatrix} -\mathbf{x}(1)^\top - \\ -\mathbf{x}(2)^\top - \\ \vdots \\ -\mathbf{x}(|\mathcal{S}|)^\top - \end{bmatrix} \qquad \mathbf{D}_\pi \doteq \begin{bmatrix} \ddots & & \mathbf{0} \\ & \mathbf{d}_\pi & \\ \mathbf{0} & & \ddots \end{bmatrix}
$$

if this "key matrix" is pos. def., then $\mathbf{A}$ is pos. def. and everything is stable

I showed in 1988 that the key matrix is pos. def. if its column sums are >0 transition prob matrix $[\mathbf{P}_\pi]_{ij} \doteq \sum_a \pi(a|i)p(j|i,a)$ where $p(j|i,a) \doteq \mathbb{P}\{S_{t+1}=j | S_t=i, A_t=a\}$

ergodic stationary distribution $[\mathbf{d}_\pi]_s \doteq d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t=s\} > 0$ $\qquad \mathbf{P}_\pi^\top \mathbf{d}_\pi = \mathbf{d}_\pi$

deterministic ‘expected’ update: $\bar{\mathbf{w}}_{t+1} \doteq (\mathbf{I} - \alpha\mathbf{A})\bar{\mathbf{w}}_t + \alpha\mathbf{b}$

Stable if $\mathbf{A}$ is positive definite
i.e., if $\mathbf{y}^\top\mathbf{A}\mathbf{y} > 0,\ \forall \mathbf{y} \neq \mathbf{0}$.
Converges to $\lim_{t\to\infty} \bar{\mathbf{w}}_t = \mathbf{A}^{-1}\mathbf{b}$.

$$\begin{aligned}
\mathbf{A} \doteq \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] &= \lim_{t\to\infty} \mathbb{E}_\pi\Big[\mathbf{x}(S_t)\left(\mathbf{x}(S_t) - \gamma\mathbf{x}(S_{t+1})\right)^\top\Big] \\
&= \sum_s d_\pi(s)\,\mathbf{x}(s)\Big(\mathbf{x}(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\mathbf{x}(s')\Big)^\top \\
&= \mathbf{X}^\top \underbrace{\mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)}\mathbf{X},
\end{aligned}$$

if this “key matrix” is pos. def., then $\mathbf{A}$ is pos. def. and everything is stable

I showed in 1988 that the key matrix is pos. def. if its column sums are >0

For the jth column, the sum is

$$\begin{aligned}
\sum_i [\mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)]_{ij} &= \sum_i\sum_k [\mathbf{D}_\pi]_{ik}[\mathbf{I} - \gamma\mathbf{P}_\pi]_{kj} \\
&= \sum_i [\mathbf{D}_\pi]_{ii}[\mathbf{I} - \gamma\mathbf{P}_\pi]_{ij} \\
&= \sum_i d_\pi(i)[\mathbf{I} - \gamma\mathbf{P}_\pi]_{ij} \\
&= [\mathbf{d}_\pi^\top(\mathbf{I} - \gamma\mathbf{P}_\pi)]_j \\
&= [\mathbf{d}_\pi^\top - \gamma\mathbf{d}_\pi^\top\mathbf{P}_\pi)]_j \\
&= [\mathbf{d}_\pi^\top - \gamma\mathbf{d}_\pi^\top]_j \\
&= (1-\gamma)d_\pi(j) \\
&> 0.
\end{aligned}$$

$$\mathbf{X} \doteq \begin{bmatrix} -\mathbf{x}(1)^\top- \\ -\mathbf{x}(2)^\top- \\ \vdots \\ -\mathbf{x}(|\mathcal{S}|)^\top- \end{bmatrix} \qquad \mathbf{D}_\pi \doteq \begin{bmatrix} \diagdown & & 0 \\ & \mathbf{d}_\pi & \\ 0 & & \diagdown \end{bmatrix}$$

2 off-policy learning problems

1. Correcting for the distribution of future returns

   solution: importance sampling (Sutton & Barto 1998, improved by Precup, Sutton & Singh, 2000), now used in GTD(λ) and GQ(λ)

2. Correcting for the state-update distribution

   solution: none known, other than more importance sampling (Precup, Sutton & Dasgupta, 2001) which as proposed was of very high variance. The ideas of that work are strikingly similar to those of emphasis…

Off-policy Temporal-Difference Learning with Linear Function Approximation

states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$

target policy $\pi(a|s)$ is no longer used to select actions

behavior policy $\mu(a|s)$ is used to select actions!

assume coverage:

$$\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$$

new ergodic stationary distribution $[\mathbf{d}_\mu]_s \doteq d_\mu(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in \mathcal{S}$

old value function $v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s] \approx \mathbf{w}_t^\top \mathbf{x}(s)$

importance sampling ratio $\rho_t \doteq \dfrac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$

$$\mathbb{E}_\mu[\rho_t|S_t = s] = \sum_a \mu(a|s)\frac{\pi(a|s)}{\mu(a|s)} = \sum_a \pi(a|s) = 1$$

For any r.v. Z_{t+1}: $\mathbb{E}_\mu[\rho_t Z_{t+1}|S_t = s] = \sum_a \mu(a|s)\dfrac{\pi(a|s)}{\mu(a|s)} Z_{t+1} = \sum_a \pi(a|s) Z_{t+1} = \mathbb{E}_\pi[Z_{t+1}|S_t = s]$

linear off-policy TD(0):

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \rho_t \,\alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t\right)\mathbf{x}_t \qquad \mathbf{x}_t \doteq \mathbf{x}(S_t)$$

$$= \mathbf{w}_t + \alpha\Big(\underbrace{\rho_t R_{t+1}\mathbf{x}_t}_{\mathbf{b}_t} - \underbrace{\rho_t \mathbf{x}_t\left(\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\right)^\top}_{\mathbf{A}_t}\mathbf{w}_t\Big)$$

and its $\mathbf{A}$ matrix:

$$\mathbf{A} = \lim_{t\to\infty}\mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty}\mathbb{E}_\mu\Big[\rho_t\mathbf{x}_t\left(\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\right)^\top\Big]$$

$$= \sum_s d_\mu(s)\mathbb{E}_\mu\Big[\rho_t\mathbf{x}_t\left(\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\right)^\top\Big|S_t = s\Big]$$

$$= \sum_s d_\mu(s)\mathbb{E}_\pi\Big[\mathbf{x}_t\left(\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\right)^\top\Big|S_t = s\Big]$$

$$= \sum_s d_\mu(s)\,\mathbf{x}(s)\bigg(\mathbf{x}(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\mathbf{x}(s')\bigg)^\top$$

$$= \mathbf{X}^\top\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{P}_\pi)\mathbf{X},$$

key matrix now has mismatched $\mathbf{D}$ and $\mathbf{P}$ matrices; it is not stable

Off-policy Temporal-Difference Learning with Linear Function Approximation

states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$

target policy $\pi(a|s)$ is no longer used to select actions
behavior policy $\mu(a|s)$ is used to select actions!

assume coverage:
$\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$

new ergodic stationary distribution $[\mathbf{d}_\mu]_s \doteq d_\mu(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in \mathcal{S}$

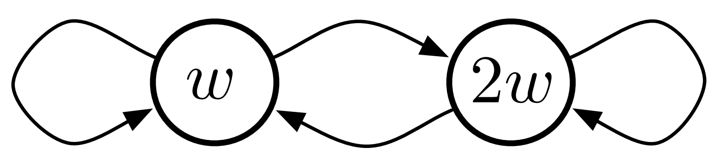
old value function $v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] \approx \mathbf{w}_t^\top \mathbf{x}(s)$

off-policy TD(0)'s $\mathbf{A}$ matrix: $\mathbf{A} = \mathbf{X}^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X},$

key matrix now has mismatched $\mathbf{D}$ and $\mathbf{P}$ matrices; it is not stable

Counterexample:

$\lambda = 0$
$\gamma = 0.9$



$\mu(\text{right}|\cdot) = 0.5$
$\pi(\text{right}|\cdot) = 1$

$\mathbf{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

transition prob matrix: $\mathbf{P}_\pi = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ $\quad [\mathbf{P}_\pi]_{ij} \doteq \sum_a \pi(a|i) p(j|i,a)$

key matrix: $\mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \times \begin{bmatrix} 1 & -0.9 \\ 0 & 0.1 \end{bmatrix} = \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix}$ sums to <0!

pos def test: $\mathbf{X}^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} -0.4 \\ 0.1 \end{bmatrix} = -0.2$

$\mathbf{A}$ is not positive definite! Stability is not assured.

2 off-policy learning problems

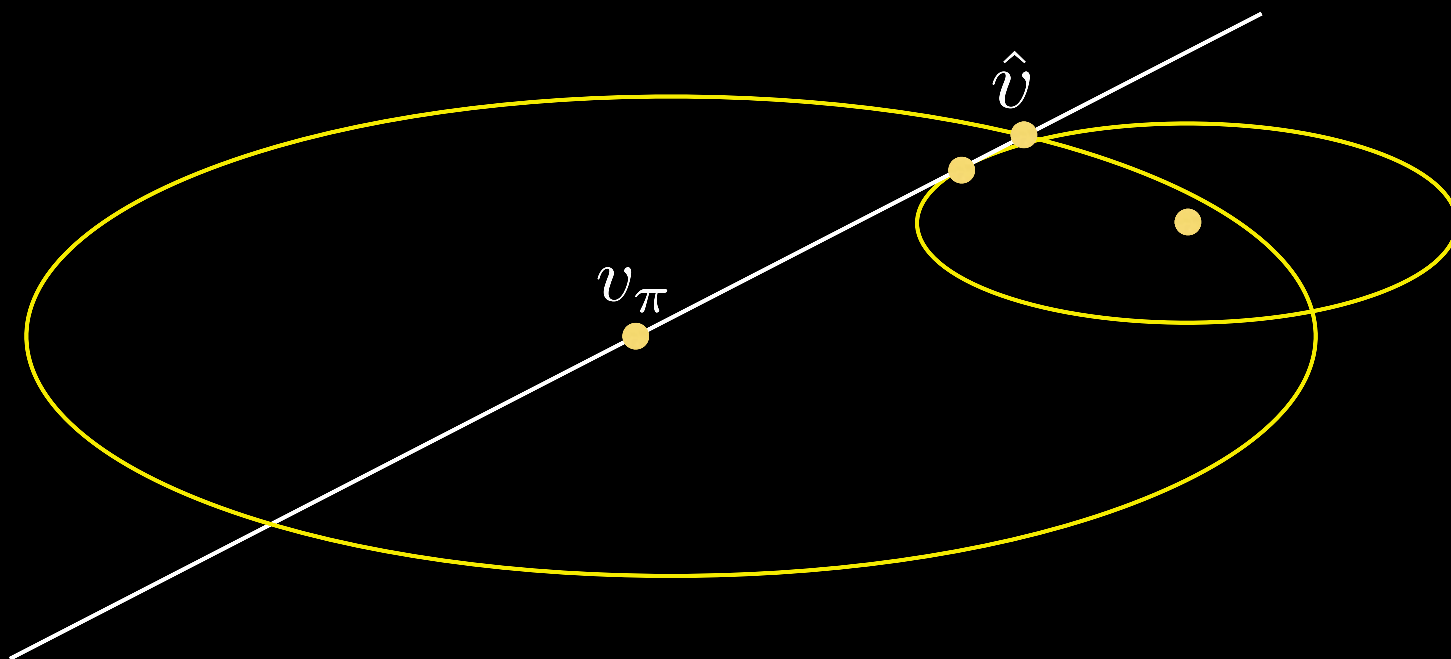
1. Correcting for the distribution of future returns

   solution: importance sampling (Sutton & Barto 1998, improved by Precup, Sutton & Singh, 2000), now used in GTD(λ) and GQ(λ)

2. Correcting for the state-update distribution

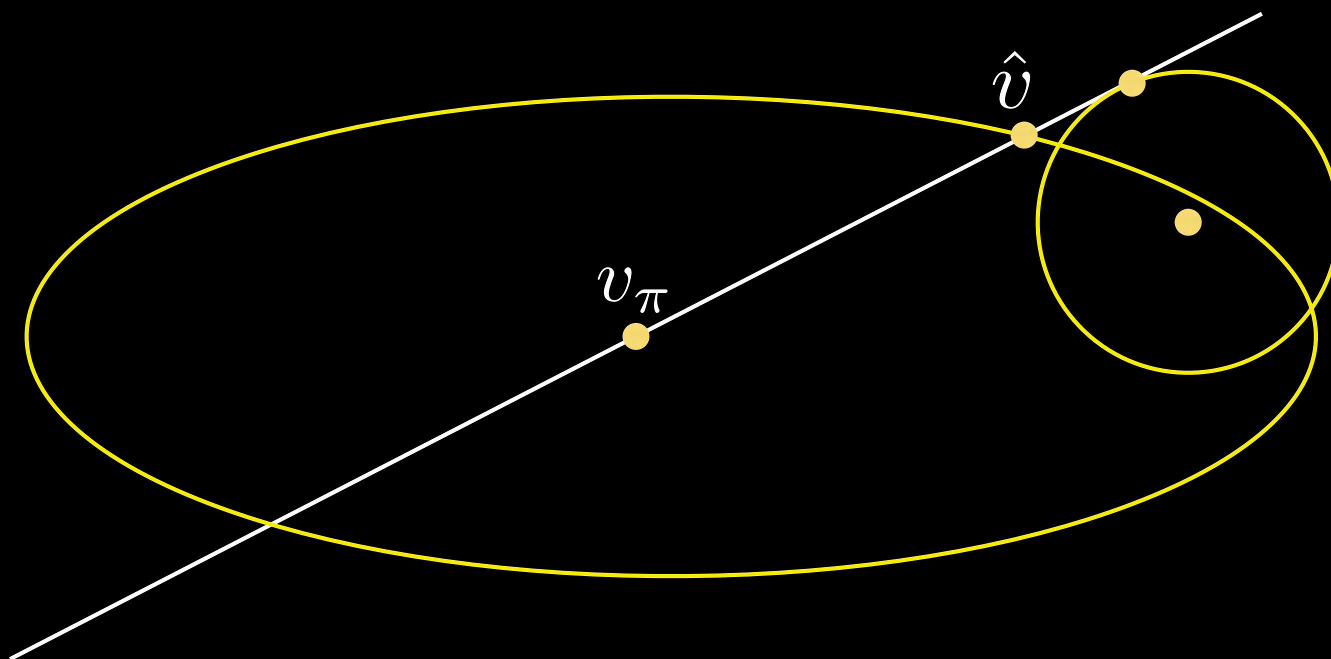
   solution: none known, other than more importance sampling (Precup, Sutton & Dasgupta, 2001) which as proposed was of very high variance. The ideas of that work are strikingly similar to those of emphasis…

Geometric Insight



Ben Van Roy 2009

Other Distribution



Problem 2 of off-policy learning: Correcting for the state-update distribution

- The distribution of updated states does not ‘match’ the target policy
- Only a problem with function approximation, but that’s a show stopper
- Precup, Sutton & Dasgupta (2001) treated the episodic case, used importance sampling to warp the state distribution from the behavior policy’s distribution to the target policy’s distribution, then did a future-reweighted update at each state
  - equivalent to emphasis = product of all i.s. ratios since the beginning of time
- ok algorithm, but severe variance problems in both theory and practice
- Performance assessed on whole episodes following the target policy
- This ‘alternate life’ view of off-policy learning was then abandoned

The *excursion* view of off-policy learning

- In which we are following a (possibly changing) behavior policy forever, and are in its stationary distribution
- We want to predict the consequences of deviating from it for a limited time with various target policies (e.g., options)
- Error is assessed on these ‘excursions’ starting from states in the behavior distribution
- Much more practical setting than ‘alternate life’
- This setting was the basis for all the work with gradient-TD and MSPBE

Emphasis warping

- The idea is that emphasis warps the distribution of updated states from the behavior policy's stationary distribution to something like the 'followon distribution' of the target policy started in the behavior policy's stationary distribution

- From which future-reweighted updates will be stable in expectation—this follows from old results (Dayan 1992, Sutton 1988) on convergence of TD(λ) in episodic MDPs

- A new algorithm: Emphatic TD(λ)

Emphatic TD(0)

Introduces a new short-term memory random variable—the *followon trace:*

$$F_t \doteq \gamma\rho_{t-1}F_{t-1} + 1, \qquad \forall t > 0 \qquad\qquad F_{-1} = 0$$

Emphatic TD(0):
$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha F_t \rho_t \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t$$
$$= \mathbf{w}_t + \alpha \Big(\underbrace{F_t \rho_t R_{t+1} \mathbf{x}_t}_{\mathbf{b}_t} - \underbrace{F_t \rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^\top}_{\mathbf{A}_t} \mathbf{w}_t \Big)$$

$$\mathbf{A} = \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t \rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^\top \right] = \mathbf{X}^\top \underbrace{\mathbf{F}(\mathbf{I} - \gamma \mathbf{P}_\pi)}_{\text{key matrix}} \mathbf{X}$$

where $\mathbf{F} \doteq \begin{bmatrix} \diagdown & & 0 \\ & \mathbf{f} & \\ 0 & & \diagdown \end{bmatrix}$

with $[\mathbf{f}]_s \doteq d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_t | S_t = s]$

we have:

$$\mathbf{f} = \mathbf{d}_\mu + \gamma \mathbf{P}_\pi^\top \mathbf{d}_\mu + \left(\gamma \mathbf{P}_\pi^\top\right)^2 \mathbf{d}_\mu + \cdots$$
$$= \left(\mathbf{I} - \gamma \mathbf{P}_\pi^\top\right)^{-1} \mathbf{d}_\mu.$$

$$\sum_i [\mathbf{F}(\mathbf{I} - \gamma \mathbf{P}_\pi)]_{ij} = \sum_i \sum_k [\mathbf{F}]_{ik} [\mathbf{I} - \gamma \mathbf{P}_\pi]_{kj}$$
$$= \sum_i [\mathbf{F}]_{ii} [\mathbf{I} - \gamma \mathbf{P}_\pi]_{ij}$$
$$= \sum_i [\mathbf{f}]_i [\mathbf{I} - \gamma \mathbf{P}_\pi]_{ij}$$
$$= [\mathbf{f}^\top (\mathbf{I} - \gamma \mathbf{P}_\pi)]_j$$
$$= [\mathbf{d}_\mu^\top (\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} (\mathbf{I} - \gamma \mathbf{P}_\pi)]_j$$
$$= [\mathbf{d}_\mu^\top]_j$$
$$= d_\mu(j)$$
$$> 0.$$

Sum of *j*th column of key matrix

Emphatic TD(0)

Introduces a new short-term memory random variable—the *followon trace:*

$$F_t \doteq \gamma \rho_{t-1} F_{t-1} + 1, \qquad \forall t > 0 \qquad\qquad F_{-1} = 0$$

Emphatic TD(0): $$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha F_t \rho_t \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t$$

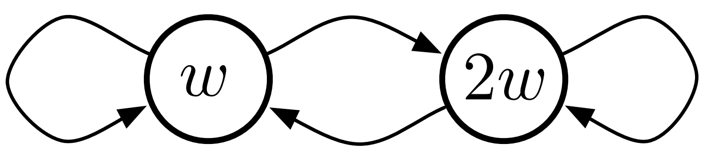
$$\mathbf{A} = \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t \rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top \right] = \mathbf{X}^\top \underbrace{\mathbf{F}(\mathbf{I} - \gamma \mathbf{P}_\pi)}_{\text{key matrix}} \mathbf{X}$$

where $\mathbf{F} \doteq \begin{bmatrix} \ddots & & \mathbf{0} \\ & \mathbf{f} & \\ \mathbf{0} & & \ddots \end{bmatrix}$

Counterexample:

$\lambda = 0$

$\gamma = 0.9$



$\mu(\text{right}|\cdot) = 0.5$

$\pi(\text{right}|\cdot) = 1$

with $[\mathbf{f}]_s \doteq d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_t | S_t = s]$

we have:

$$\mathbf{f} = \mathbf{d}_\mu + \gamma \mathbf{P}_\pi^\top \mathbf{d}_\mu + \left(\gamma \mathbf{P}_\pi^\top \right)^2 \mathbf{d}_\mu + \cdots$$
$$= \left(\mathbf{I} - \gamma \mathbf{P}_\pi^\top \right)^{-1} \mathbf{d}_\mu.$$

$$[\mathbf{f}]_1 = d_\mu(1) = 0.5$$

$$[\mathbf{f}]_2 = 0.5 + 0.9 + 0.9^2 + 0.9^3 + \cdots$$
$$= 0.5 + 0.9 \cdot 10$$
$$= 9.5$$

$$\mathbf{P}_\pi = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{F}(\mathbf{I} - \gamma \mathbf{P}_\pi) = \underbrace{\begin{bmatrix} 0.5 & 0 \\ 0 & 9.5 \end{bmatrix}}_{\mathbf{F}} \times \underbrace{\begin{bmatrix} 1 & -0.9 \\ 0 & 0.1 \end{bmatrix}}_{\mathbf{I} - \gamma \mathbf{P}_\pi} = \underbrace{\begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.95 \end{bmatrix}}_{\text{key matrix}}$$

sums to >0

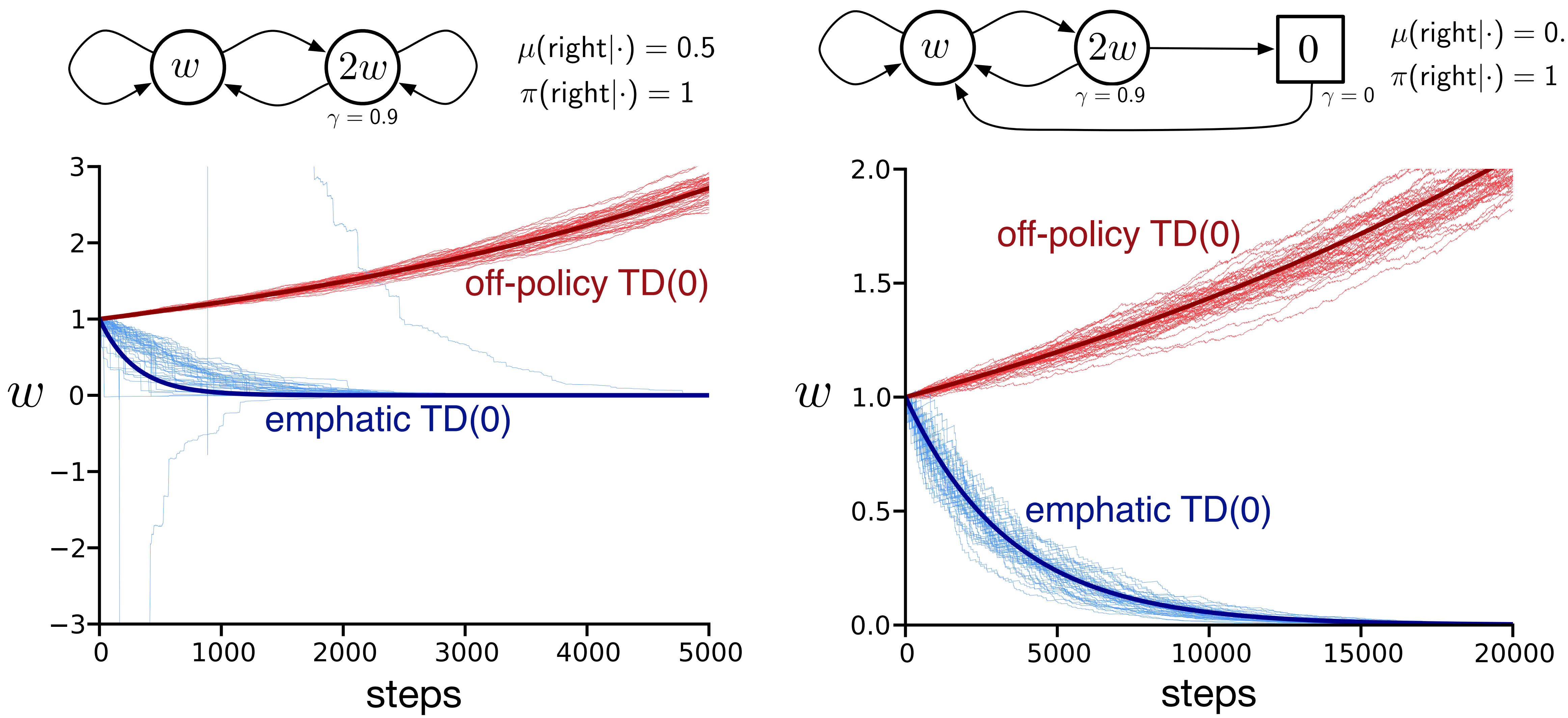


Figure 3: Emphatic TD approaches the correct value of zero, whereas conventional off-policy TD diverges, on fifty trajectories on the $w \to 2w$ problems shown above each graph. Also shown as a thick line is the trajectory of the deterministic expected-update algorithm. On the continuing problem (left) emphatic TD has occasional high variance deviations from zero.

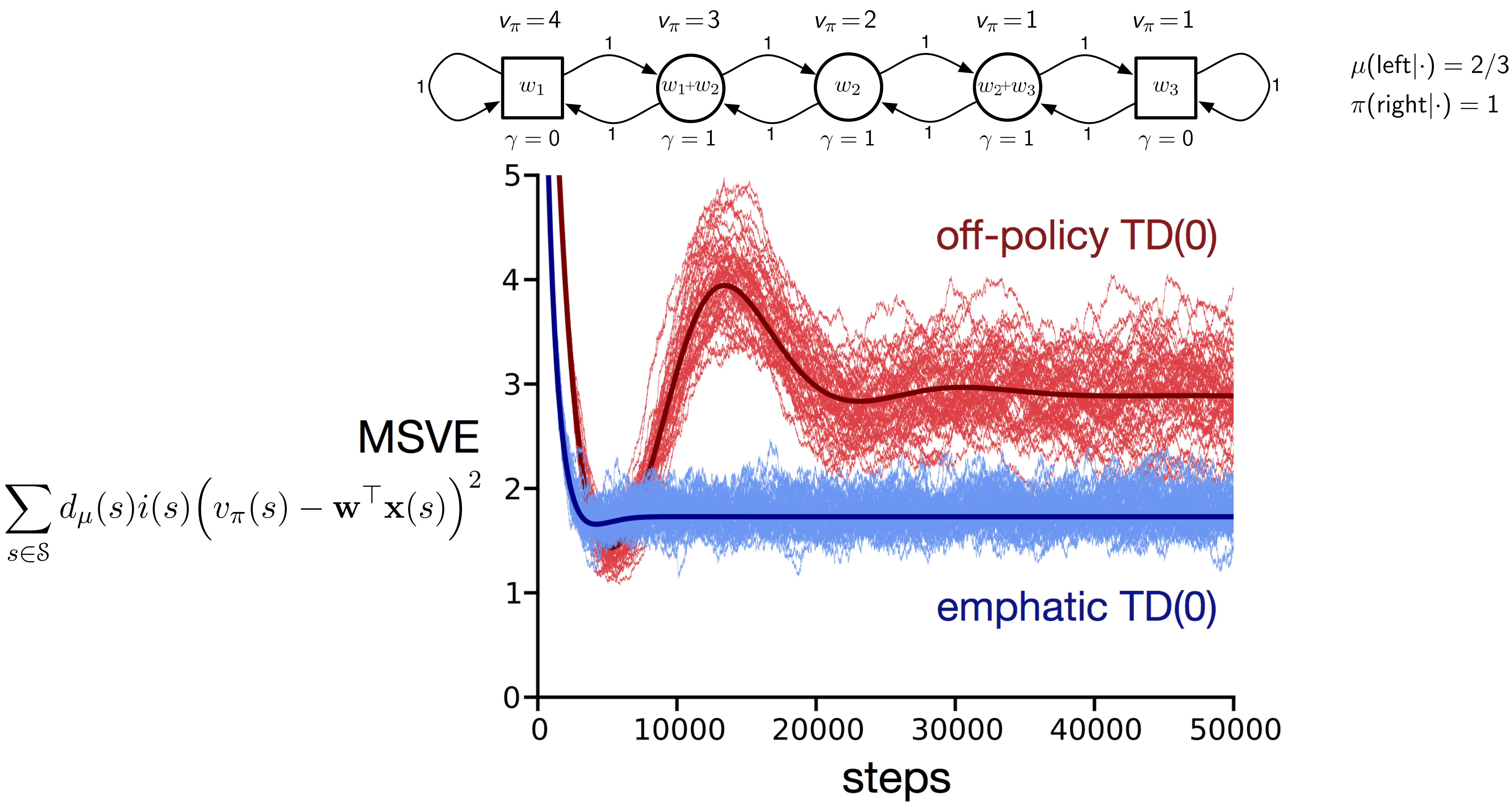


Figure 4: Twenty learning curves and their analytic expectation on the 5-state problem from Section 5, in which excursions terminate promptly and both algorithms converge reliably. Here $\lambda = 0$, $\mathbf{w}_0 = \mathbf{0}$, $\alpha = 0.001$, and $i(s) = 1, \forall s$. The MSVE performance measure is defined in (20).

Summary of emphatic results

- Linear emphatic TD(0) is the simplest TD alg with linear FA that is stable under off-policy training
- Some empirical illustrations
- Stability theorem for full case of GVFs
- Convergence w.p.1 theorem (Janey Yu, under review)
- Asymptotic approximation bounds (Remi Munos)
- Also a new (better?) algorithm for the on-policy case