

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 parameter 2 [0*,* 1). Conventional linear TD(0) is defined by the following update to the parameter 2 [0*,* 1). Conventional linear TD(0) is defined by the following update to the parameter vector ^w*^t* ² ^R*n*, made at each of a sequence of time steps *^t* = 0*,* ¹*,* ²*,...*, on Tomporal Difforongo Loorning with Lingar Eunotian Approvimation **Temporal-Difference Learning with Linear Function Approximation**
 C^S actions $A_i \subseteq A$ rowards $B_{i,i} \subset \mathbb{R}$ policy $\pi(a|s) = \mathbb{R}[A_i - a|S_i - s]$ α - Oifference Learning with Linear Function Annroximation **Temporal-Difference Learning with Linear Function Approximation**
 $\sum_{n=0}^{\infty} a_n = a_n$

 $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$
 \downarrow [**p**].. \div $\sum_{\pi(a|i) n(i|i-a) \leq \pi}$ $\lim_{t \to \infty} \frac{1}{n} \int_{0}^{\infty} \frac{1}{t^{2}} e^{-t} dt = \lim_{t \to \infty} \frac{1}{t^{2}} \left[\frac{1}{t^{2}} \frac{1}{t^{2}} + \frac{1}{t^{2}} \frac{1}{t^{2}} \frac{1}{t^{2}} \right] = \lim_{t \to \infty} \frac{1}{t^{2}} \left[\frac{1}{t^{2}} \frac{1}{t^{2}} + \frac{1}{t^{2}} \frac{1}{t^{2}} \frac{1}{t^{2}} \frac{1}{t^{2}} \right] = \lim_{t \to \infty} \frac{1}{t^{2}} \left[\frac{$ atrix $[\mathbf{F}_{\pi}]_{ij} = \sum_{a} \pi(a|i) p(j|i, a)$ where $p(j|i, a) = \mathbb{P}\{\mathcal{D}_{t+1} = j | \mathcal{D}_{t} = i, A_{t} = a\}$ v distribution $[\mathbf{d}_\pi]_s \doteq d_\pi(s) \doteq \lim_{t \to \infty} \mathbb{P}\{S_t\!=\!s\} > 0$. P $_\pi^\top \mathbf{d}_\pi = \mathbf{d}_\pi$ $\gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$ $0 \leq \gamma < 1$ feature vectors $\mathbf{x}(s) \in \mathbb{R}^n$ $\forall s \in \mathcal{S}$ \cdot positive semi-definite. $\mathbb{E}[\mathbb{E}[E] \sim \mathbb{E}[\mathbb{E}$ $\pi_1 \in \mathbb{R}$ policy $\pi(a|s) = \mathbb{P}\{A_t = a|S_t = s\}$ $D \leq \gamma < 1$ feature vectors $\mathbf{x}(s) \in \mathbb{R}^n$ $\forall s \in \mathcal{S}$ value function $v_{\pi}(s) = \mathbb{E}_{\pi}[G_t | S_t = s] \approx \mathbf{w}_t \cdot \mathbf{x}(s)$ weight vector $\mathbf{w}_t \in \mathbb{R}^n \cdot n \ll |\delta|$
weight vector $\mathbf{w}_t \in \mathbb{R}^n \cdot n \ll |\delta|$ states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ policy $\pi(a|s) = \mathbb{P}\{A_t = a|S_t = s\}$ ergodic stationary distribution $[\mathbf{a}_{\pi}]_s = a_{\pi}(s) = \lim_{t\to\infty} \mathbb{P}\{\beta_t = s\} > 0$
return $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = 0 \leq \gamma < 1$ feature vecto distribution $[\mathbf{d}_\pi]_s = d_\pi(s) = \lim_{t \to \infty} \mathbb{P}$ x(*St*)*,* (1) $\textsf{linear TD} (0): \ \mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x} (S_{t+1}) - \mathbf{w}_t^\top \mathbf{x} (S_t) \right)$ value function $v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t\!=\!s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ dialishion prop matrix $[\mathbf{I}_{\pi}]_{ij} = \sum_{a} \pi \langle a | i \rangle p \langle j | i, a \rangle$ where $p(j | i, a) = \mathbb{I}_{\{\mathcal{D}_{t+1} = j | \mathcal{D}_{t} = i, A_t = a\}}$
ergodic stationary distribution $[d_{\pi}]_a \doteq d_{\pi}(s) \doteq \lim_{\epsilon \to a} \mathbb{P}\{S_t = s\} > 0$ $\mathcal{U}_{t+3} + \cdots \hspace{.5cm} 0 \leq \gamma < 1$ feature vectors $\mathbf{x}(s) \in \mathbb{R}^n$ $\forall s \in \mathcal{S}$ $\mathbf{s}(s)$ is a statute vectors $\mathbf{x}(s) \in \mathbb{R}^n$ $n \ll |\mathcal{S}|$ $f_{\mathbf{r}} + \alpha \left(R_{t+1} + \alpha \mathbf{w}^{\top} \mathbf{x} (S_{t+1}) - \mathbf{w}^{\top} \mathbf{x} (S_t) \right) \mathbf{x} (S_t)$ transition prob matrix $[\mathbf{P}_{\pi}]_{ij} = \sum_{a} \pi(a|i)p(j|i,a)$ where $p(j|i,a) = \mathbb{P}\{S_{t+1} = j|S_t = i, A_t = a\}$ value function $v_{\pi}(s) = \mathbb{E}_{\pi}[G_t|S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ weight vector $\mathbf{w}_t \in \mathbb{R}^n$ $n \ll |\mathcal{S}|$
 $\left(\sum_{i=1}^n v_i^{\top} S_i\right)$ $\mathbf{H}_{\pi} = \mathbf{H}_{\pi}(v) = \min_{t \to \infty} \mathbf{H}_{\pi}(v_t - v)$ $v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t|S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ we of learned parameters, *[|]*S*[|] .* $\begin{array}{ccc} \text{m} \text{norm} & \text{m} \text{norm} \\ \text{m} \text{norm} & \text{m} \text{norm} \end{array}$ transition prop matrix $[\mathbf{F}_{\pi}]_{ij} = \sum_{a} \pi(a|i)p(j|i,a)$ where $p(j|i,a) = \mathbb{E} \{ \partial t + 1 - j | \partial t - i, A_t - a \}$
ergodic stationary distribution $[\mathbf{d}_{\pi}]_s \doteq d_{\pi}(s) \doteq \lim_{t \to \infty} \mathbb{P} \{ S_t = s \} > 0$ $\mathbf{P}_{\pi}^{\top} \mathbf{d}_{\pi} = \mathbf{d}_{\pi}$ return $G_t \doteq B_{t+1} + \gamma B_{t+2} + \gamma^2 B_{t+2} + \cdots$ $0 < \gamma < 1$ $\begin{array}{ccc} \overline{u} & \overline{v} & \overline$ transition prob matrix $[\mathbf{P}_{\pi}]_{ij} = \sum_a \pi(a|i)p(j|i,a)$ where $p(j|i,a) = \mathbb{P}\{S_{t+1} = j|S_t = i, A_t = a\}$ $\mathbf{E}_t | S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ weight vector $\mathbf{w}_t \in \mathbb{R}^n$ $n \ll |\mathcal{S}|$ $\mathcal{A} \subset \mathbb{P}$ policy $\pi(a|s) - \mathbb{P}[A-a|S-a]$ $\tau_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ policy $\pi(a|s) = \mathbb{P}\{A_t = a|S_t = s\}$ inear $TD(0)$: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t$ transition prob matrix $[\mathbf{P}_{\pi}]_{ij} = \sum_{a} \pi(a|i)p(j|i,a)$ where $p(j|i,a) = \mathbb{P}\{S_{t+1} = j|S_t = i, A_t = a\}$ $\gamma < 1$ $\begin{array}{lll} \mathbf{F} & \math$ $\begin{array}{ccc} \cdot & 0 \leq \gamma < 1 & \text{feature vectors} \ \mathbf{x}(s) \in \mathbb{R}^n & \forall s \in \mathcal{S} \end{array}$ ⇣ *^t* x(*St*) $v^2 R_{t+3} + \cdots \quad 0 \le \gamma < 1$ feature vectors $\mathbf{x}(s) \in \mathbb{R}^n \ \ \forall s \in \mathcal{S}$ $s=s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ weight vector $\mathbf{w}_t \in \mathbb{R}^n$ $n \ll |\mathcal{S}|$ $f: \mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}(S_{t+1}) - \mathbf{w}_t^\top \mathbf{x}(S_t) \right) \mathbf{x}(S_t)$ weight vector $\mathbf{w}_t \in \mathbb{R}^n$ $n \ll |\mathcal{S}|$ \overline{C} \mathbb{D} in equality by definition rather than \overline{C} \mathbb{D} in that follows \overline{C} $p(i|i, a) \doteq \mathbb{P}\{S_{t+1} = i | S_t = i, A_t = a\}$ function approximation \mathbf{F} and \mathbf{F} $\mathbf{w}_t^\top \mathbf{x}(S_{t+1}) - \mathbf{w}_t^\top \mathbf{x}(S_t) \Big) \mathbf{x}(S_t)$ $n \ll |\mathcal{S}|$ from the previous definition of the action of $u(v) = \frac{u}{r} \left(2r - u|v_t - v \right)$
the set of actions and the above $v(i|i, \alpha)$ in $\mathbb{D}\left[C_i - i|S_i - i|A_i - \alpha\right]$ $\rho(x | \mathbf{P}_{\pi}|_{ij} \doteq \sum_{a} \pi(a|i) p(j|i,a)$ where $p(j|i,a)$
 $\rho(x | i)$ distribution $\mathbf{d}_{\pi}|_{s} \doteq d_{\pi}(s) \doteq \lim_{\epsilon \to \infty} \mathbb{P}\{S_{t} = s\} > 0$ are assumed to be finited to be finitely but the number of states is assumed much larger than the number of $\mathbf{u}_\pi \mathbf{u}_\pi - \mathbf{u}_\pi$ $-\gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \qquad 0 \leq \gamma < 1$
 $\zeta_{\mathcal{S}} = \mathbb{E} \left[C \log |S| - \epsilon \right] \approx \mathbf{W}^\top \mathbf{Y}(\epsilon)$ $\mathbf{x} \in \mathbb{R}^n \quad \forall s \in \mathbb{S}$
neight vector $\mathbf{x}(s) \in \mathbb{R}^n \quad \forall s \in \mathbb{S}$ $\begin{aligned} \mathcal{L}(S) = \mathbb{E}_{\pi}[\mathbf{G}_t|\mathcal{S}_t = S] \approx \mathbf{W}_t \mathbf{X}(S) \quad &\text{Welght Vector} \quad \mathbf{W}_t \in \mathbb{R}^{n \times |I|} \ll |\mathcal{O}| \end{aligned}$ states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ policy $\pi(a|s) = \mathbb{P}\{A_t = a|S_t = s\}$ t_{tot} attend and initial transient, states will be visited and t_{tot} and uansition prob matrix $[\mathbf{F}_{\pi}]_{ij} = \sum_a \pi(a|i)p(j|i,a)$ where $p(j|i,a) = \mathbb{I}$ ergodic stationary distribution $[\mathbf{d}_\pi]_s \doteq d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t\!=\!s\} > 0$ $\mathbf{P}_\pi^\top \mathbf{d}_\pi = \mathbf{d}_\pi$ return $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$ $0 \leq \gamma < 1$ foature vectors $\mathbf{x}(s) \in \mathbb{R}^n$ $\forall s \in \mathcal{S}$ c_{n+1} , c_{n+2} , c_{n+3} , c_{n+2} , c_{n+3} , c_{n+1} , c_{n+2} , c_{n+3} , c_{n+1} , c_{n+1} , c_{n+2} , c_{n+3} , c_{n+1} , c_{n+1} , c_{n+1} , c_{n+2} , c_{n+1} , c_{n+1} , c_{n+1} , c_{n+1} , c_{n+1} , c_{n+1} , value function $v_{\pi}(s) = \mathbb{E}_{\pi}[\mathbf{G}_t | \mathcal{S}_t = s] \approx \mathbf{W}_t[\mathbf{X}(s)]$ weight vectors linear TD(0): $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x} (S_{t+1}) \right)$ $\arg\limits_{t\in [0,n]}\mathcal{E}_{t+1}\in\mathbb{R}$ politically $\pi(a|s) = \mathbb{P}\{A_t = a|S_t = s\}$
 $\pi(i|i, a) \doteq \mathbb{P}\{S_{t+1} = i|S_t = i|A_t = a\}$ $\left[\frac{n}{2} \right]$ $\mu(\nu, \mu(\nu, \omega))$ where $\mu(\nu, \mu) = \mu$ $\mathsf{ctor} \leftarrow \mathbf{w}_t \in \mathbb{R}^n \enspace n \ll |\mathcal{S}|$ th data from \overline{A} continuing \overline{D} decision $\pi(a|a)$ in $\mathbb{D}[A] = a|\overline{C} = a$ function approximation case, that of linear TD() with = 0 and constant discount-rate parameter $\sum_a \mu(a|b)p(j|b,a)$ where $p(j|b,a) = \mu(b|b+1-j)b(t-b, At-a)$ $R_{\tau}(s) \doteq \mathbb{F}_{\tau}$ $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha$ $\sqrt{2}$ \setminus

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

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

1. A real matrix A is definite to be *positive definition* in the *positive definition* is stable. Column sums are >0.

 \mathbb{R}

 \mathcal{L} \mathcal{L} is statistic section in the statistic

stable

^t ^x(*s*) ⇡ *^v*⇡(*s*) *.*

= E⇡[*Gt|S^t* =*s*] *,* (2)

 \int $\frac{1}{2}$ \int $\frac{1}{$

n are $>$ 0

$$
= \mathbf{w}_t + \alpha (\mathbf{b}_t - \mathbf{A}_t \mathbf{w}_t)
$$

$$
= (\mathbf{I} - \alpha \mathbf{A}_t) \mathbf{w}_t + \alpha \mathbf{b}_t.
$$

 $N = \frac{1}{2}$ return to $\left(\mathbf{T} - \mathbf{A}\right) = \frac{1}{2}$. . deterministic 'expected' update: $\bar{\mathbf{w}}_{t+1} \doteq (\mathbf{I} - \alpha \mathbf{A}) \bar{\mathbf{w}}_t + \alpha \mathbf{b}$ $\mathbf{u}_t + \mathbf{u}_t = \mathbf{v}$ w_t**+**1 = w^t₁x(*S*_t) = wt_i² = wti₁x(*S*^t) = wti₁ istic expe λ (\mathbf{F} ed update: ${\mathbf w}_{t+1} = ({\mathbf \pmb{\bot}}-1)$ A*t*2R*n*⇥*ⁿ* whected' undar $\overline{W}_{t+1} =$ $\begin{array}{c} \begin{array}{c} \circ \\ \circ \\ \circ \end{array} \end{array}$ $(\mathbf{I} - \alpha \mathbf{A}) \mathbf{\bar{w}}_t + \alpha \mathbf{b}$ π $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $\texttt{update:} \;\; \bar{\mathbf{w}}_{t+1}$ = $\overline{}$ deterministic 'expected' update: $\bar{\mathbf{w}}_{t+1} \doteq (\mathbf{I} - \alpha \mathbf{A}) \bar{\mathbf{w}}_t + \alpha \mathbf{b}$ from previous definitions. In on-policy training, the actions are chosen according to a target deterministic 'expected' update: $\overline{\mathbf{w}}_{t+1} \doteq (\mathbf{I} - \alpha \mathbf{A}) \overline{\mathbf{w}}_t + \alpha \mathbf{b}$ update apade are \mathbf{u}_{t+1} (* \mathbf{u}_{t+1} and the set of \mathbf{u}_{t+1} pected' update: $\bar{\mathbf{w}}_{t+1} \doteq (\mathbf{I} - \alpha \mathbf{A}) \bar{\mathbf{w}}_t + \alpha \mathbf{b}$ are assumed to be finite, but the number of states is assumed much larger than the number of states is assumed much larger than the number of states is assumed much larger than the number of states is assumed much larger

Rt+1x(*St*)

1. A real matrix A is defined to be *positive definite* in this paper i↵ y>Ay *>* 0 for any vector y 6= 0.

 L all diagonal everythic

with the window of the win

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

\nStable if **A** is positive definite $\mathbf{A} = \lim_{t \to \infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t \to \infty} \mathbb{E}_{\pi} \Big[\mathbf{x}(S_t) \left(\mathbf{x}(S_t) - \gamma \mathbf{x}(S_{t+1}) \right)^{\top} \Big]$
\ni.e., if $\mathbf{y}^{\top} \mathbf{A} \mathbf{y} > 0$, $\forall \mathbf{y} \neq \mathbf{0}$.
\nConverges to $\lim_{t \to \infty} \bar{\mathbf{w}}_t = \mathbf{A}^{-1} \mathbf{b}$.
\n
$$
= \sum_s d_{\pi}(s) \mathbf{x}(s) \Big(\mathbf{x}(s) - \gamma \sum_{s'} [\mathbf{P}_{\pi}]_{ss'} \mathbf{x}(s') \Big)
$$
\n
$$
= \mathbf{X}^{\top} \mathbf{D}_{\pi} (\mathbf{I} - \gamma \mathbf{P}_{\pi}) \mathbf{X},
$$
\n
$$
\mathbf{X} = \begin{bmatrix} -\mathbf{x}(1)^{\top} - \\ -\mathbf{x}(2)^{\top} - \\ \vdots \\ -\mathbf{x}(|\mathcal{S}|)^{\top} - \end{bmatrix} \quad \mathbf{D}_{\pi} = \begin{bmatrix} \mathbf{d}_{\pi} \\ \mathbf{0} \end{bmatrix}
$$
\nif this "key matrix" I showed in 1988 is pos. def., then that the key matrix **A** is pos. def. and is pos. def. if its everything is stable column sums are >0

x(*St*) (x(*St*) x(*St*+1))>

 α $[\mathbf{P}_{\pi}]_{ii} \doteq \sum \pi(a|i)p(i|i,a)$ where $p(i|i,a) \doteq \mathbb{P}\{S_{t+1} = i|S_t = i, A_t = a\}$ scope of this paper. Here we consider only the positive definitely definitely \sim A, which we consider \sim A, which SUIDUIIOII $[\mathbf{u}_{\pi}]_s = u_{\pi}(s) = \min_{t\to\infty} \mathbb{P}\{\partial_t = s\} > 0$ that once the process is in it, it remains in it. Let P⇡ denote the *N* ⇥*N* matrix of transition transition prob matrix $[\mathbf{P}_{\pi}]_{ij} = \sum_a \pi(a|i)p(j|i,a)$ where $p(j|i,a) = \mathbb{P}\{S_{t+1} = j|S_t = i, A_t = a\}$ tationary distribution $[d^{\top}] \doteq d^{\top}(s)$ $\tau \sim \pi$ a_n *tropolition arab motrix* $[D]$, $\therefore \sum_{\pi} (q|i) p(i|i|q)$ where $p(i|i|q)$ \therefore $D[C]$, \therefore $d = q$ ergodic stationary distribution $[\mathbf{d}_\pi]_s \doteq d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t\!=\!s\} > 0$ $\mathbf{P}_\pi^\top \mathbf{d}_\pi = \mathbf{d}_\pi$ under probarration $\left[\pm \pi y\right]$ $\left(\pm a^n (u|v)p(y|v,w) \right)$ where $p(y|v,w)$ $a(i)p(i|i, a)$ where $p(i|i, a) \doteq P{S_{t+1} = i|S_t = i, A_t = a}$ a [}] $\mathbf{f}(\mathbf{y}|\mathbf{y}, \mathbf{z})$ is the special property of $\mathbf{f}(\mathbf{y}|\mathbf{y}, \mathbf{z})$ F a consider the interpretation of $d \cdot d$ (a) is $f(x) = f(x) - 1$ orgoalo dialionaly didindulare $[\alpha_{\pi}]_s$ w_{π}(e) $\min_{l\to\infty}$ μ (e) ℓ

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

\nStable if **A** is positive definite
\ni.e., if $\mathbf{y}^\top \mathbf{A} \mathbf{y} > 0$, $\forall \mathbf{y} \neq \mathbf{0}$.
\nConverges to $\lim_{t \to \infty} \vec{w}_t = \mathbf{A}^{-1} \mathbf{b}$.
\n
$$
= \sum_s d_\pi(s) \mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \sum_{s'} [\mathbf{P}_\pi]_{ss'} \mathbf{x}(s') \right)^\top
$$
\n
$$
= \mathbf{x}^\top \mathbf{D}_\pi (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X}
$$
\nFor the *j*th column, the sum is
\n
$$
\sum_i [\mathbf{D}_\pi (\mathbf{I} - \gamma \mathbf{P}_\pi)]_{ij} = \sum_i \sum_k [\mathbf{D}_\pi]_{ik} [\mathbf{I} - \gamma \mathbf{P}_\pi]_{kj}
$$
if this "key matrix" I showed in 1988
\n
$$
= \sum_i [\mathbf{D}_\pi (\mathbf{I} - \gamma \mathbf{P}_\pi)]_{ij}
$$
\n
$$
= \sum_i [D_\pi]_{ii} [\mathbf{I} - \gamma \mathbf{P}_\pi]_{ij}
$$
\n
$$
= [\mathbf{d}_\pi^\top (\mathbf{I} - \gamma \mathbf{P}_\pi)]_j
$$
\n
$$
= [\mathbf{d}_\pi^\top - \gamma \mathbf{d}_\pi^\top]_{ij}
$$
\n
$$
= \mathbf{d}_\pi(\mathbf{i}) \mathbf{I} - \mathbf{i} \mathbf{j} \mathbf{k}
$$
\n
$$
\mathbf{X} = \begin{bmatrix} -\mathbf{x}(1)^\top - \\ -\mathbf{x}(2)^\top - \\ -\
$$

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 $GTD(\lambda)$ and $GQ(\lambda)$

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 $R_{t+1} \in \mathbb{R}$ $P^{\text{tr}(\alpha|S)}$ is used to select actions:
bw ergodic stationary distribution $[\mathbf{d}_{\mu}]_s \doteq d_{\mu}(s) \doteq \mathrm{Id}_s$ $\mu(s)$ is used new ergodic stationary distribution $[\mathbf{d}_{\mu}]_s \doteq d_{\mu}(s) \doteq \lim_{t \to \infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in \mathcal{S}$ ω value function $v_{\pi}(s) = \frac{1}{\pi} \left[\frac{U_t}{V_t - s} \right] \approx \mathbf{w}_t \mathbf{A}(s)$ $\pi(A | S_t)$ $\pi(a | s)$ importance sampling ratio $\rho_t \doteq \frac{\pi(A_t|S_t)}{\pi(A \perp S)}$ $\mathbb{E}_{\mu}[\rho_t|S_t = s] = \sum \mu(a|s) \frac{\pi(a|s)}{\pi(s)} = \sum$ $f(x) = \mu(u|s)$ the actions are chosen assumed as the actions are chosen For any r.v. Z_{t+1} : $\mathbb{E}_{\mu}[\rho_t Z_{t+1} | S_t = s] = \sum_{\mu} \mu(a|s) \frac{\pi(a|s)}{\mu(a|s)} Z_{t+1} = \sum_{\pi} \pi(a|s) Z_{t+1} = \mathbb{E}_{\pi}[Z_t]$ $a \longrightarrow \mu(u|\omega)$ assumed much states is assumed much larger than the number of states is assumed much larger than the number of a $\textsf{linear off-policy TD(0):} \quad \mathbf{w}_{t+1} \doteq \mathbf{w}_t + \rho_t \, \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t \qquad \qquad \mathbf{x}_t$ $\mathbf{u} = \mathbf{w} \cdot \mathbf{v} \cdot \mathbf{v$ $\frac{1}{2}$ \overline{a} | \overline{a} | \overline{a} | \overline{a} | \overline{a} | and its **A** matrix: $\mathbf{A} = \lim_{t \to \infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t \to \infty} \mathbb{E}_{\mu} \left[\rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top \right]$ $-\sum_{s} a_{\mu}(s) \mathbb{E}_{\mu} \left[\rho_t \mathbf{x}_t \mathbf{x}_t - \gamma \mathbf{x}_{t+1} \mathbf{y} \mid \mathcal{O}_t - s \right]$ ⁼ *^Rt*+1 ⁺ *Rt*+2 ⁺ 2*Rt*+3 ⁺ *··· .* (3) Γ and Γ matrices the stability is the stabili x(*St*) (x(*St*) x(*St*+1))> $\text{UQ Value IURUOUT } v_{\pi}(s) = \mathbb{E}_{\pi}[\mathbf{G}t|\mathcal{D}t = s] \approx \mathbf{W}_t \mathbf{X}(s)$ importance sampling ratio $\rho_t = \frac{\pi (A_t | \mathcal{S}_t)}{\sigma_t | \mathcal{S}_t}$ $\mathbb{E}_u[\rho_t | \mathcal{S}_t]$ $\mu(\mathcal{A}t|\mathcal{O}t)$ For any r.v. Z_{t+1} : $\mathbb{E}_{\mu}[\rho_t Z_{t+1} | S_t = s] = \sum_{t} \mu(a|s) \frac{\pi(a|s)}{\mu(a|s)} Z_t$ \overline{a} assumed much have number of $\mu(a|s)$ $\begin{array}{c} \n\cdot \\
\cdot \\
\cdot\n\end{array}$ $= \mathbf{W}_t + \alpha \left(\underbrace{\rho_t \mathbf{K}_{t+1} \mathbf{X}_t}_{t+1} - \underbrace{\rho_t \mathbf{X}_t (\mathbf{X}_t - \gamma \mathbf{X}_{t+1})}_{t+1} \mathbf{W}_t \right)$ α *x* k ^{*R*} k ^{*R*} k ^{*R*} k ^{*R*} k ^{*R*} k ^{*R*} k ^{*R*} k ^{*R*} k ^{*R* k} k *R* k $\sum_{\mu} a_{\mu}(s) \mathbf{x}(s) \mathbf{x}(s) \mathbf{x}(s) = \gamma \sum_{s'} [\mathbf{F} \pi] s$ $\frac{1}{2}$ w*^t* $\pi(a|s) > 0 \implies \mu(a|s) > 0$ ⇣ assume coverage: ⇡(*At|St*) $\forall s,a$ $u_{\mu}(s) = \min_{t \to \infty} \mathbb{E}\{\partial_t = s\} > 0, \forall s \in \mathcal{S}$ old value function $v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t\!=\!s] \approx \mathbf{w}_t^\top \mathbf{x}(s)$ from previous definitions. In order $\pi(a|s)$ the action actions are chosen actions are chosen actions are chosen actions and $\mathbb{E}_{\mu}[\rho_t|S_t\!=\!s] = \sum_{a}\mu(a|s)\frac{\mu(a|s)}{\mu(a|s)} = \sum_{a}\pi(a|s) = 1$ $\pi(a|s)$ $\chi(a|s) \frac{\pi(a|s)}{\mu(a|s)} Z_{t+1} = \sum_{a} \pi(a|s) Z_{t+1} = \mathbb{E}_{\pi}[Z_{t+1}|S_t\!=\!s].$ $\begin{array}{ccc} & & \alpha & \\ \tau & & \tau & \end{array}$ $\mathbf{x}_{t+1} + \gamma \mathbf{w}_t \mathbf{x}_{t+1} - \mathbf{w}_t \mathbf{x}_t \mathbf{x}_t$ $\mathbf{x}_t = \mathbf{x}(\mathcal{A}_t)$ $\mathbf{E} = \sum d_{\mu}(s) \mathbb{E}_{\mu} \left[\rho_t \mathbf{x}_t (\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^\top \middle| S_t = s \right]$ *^G^t .* ⁼ *^Rt*+1 ⁺ *Rt*+2 ⁺ 2*Rt*+3 ⁺ *··· .* (3) The TD(0) update (1) can be rewritten to make the stability issues more transparent: s' $\mathbf{x}(s)$ $\frac{1}{2}$ target policy $\pi(a|s)$ is no longer used to select actions assume coverage: linear off-policy TD(0): $in point$ is important binoronoo Lourining with</u> states $D_t \in \mathcal{O}$ actions $\Lambda_t \subset \Lambda$ rewards $n_{t+1} \subset \mathbb{R}$ behavior policy $\mu(a|s)$ is used to select actions! $\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$ $\text{new ergodic stationary distribution } [d_u]_s = d_u(s) = \lim_{t \to \infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in S$ $f_{\text{reconstruction}}$ from the target policy. For example, some action $\alpha - \pi(A_t|S_t)$ importance sampling ratio $r \iota$ $\mu(A_t|S_t)$ $\mu(\mu_t|\nu)$ τ action action action actually $\pi(a|s)$ ⇡(*At|St*) *µ*(*At|St*) $\mathbf{D}(\mathbf{0})$: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \rho_t \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right)$ s **A** matrix: $\mathbf{A} = \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu \left[\rho_t \mathbf{x}_t \left(\mathbf{x}_t \right) \right]$ $\frac{s}{\Box}$ $\mathbf{p}=\sum d_{\mu}(s)\mathbb{E}_{\pi}\big|\mathbf{x}_{t}\left(\mathbf{x}_{t}-\gamma\mathbf{x}_{t+1}\right)^{\top}\big|S_{t}=s\big|_{\mathcal{S}_{t}}$ $\bm s$ the action taken was more or less likely matrix now **Off-policy** Temporal-Difference Learning with Linear Function Approximation $\mathcal{A}_t \in \mathcal{A}$ in rewards $R_{t+1} \in \mathbb{R}$ $i \in \mathcal{U}$ is ignored the values that $i \in \mathcal{U}$ if the action state $i \in \mathcal{U}$ \mathcal{A}_μ (*s*) $\dot{=}$ $\lim_{t\to\infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in \mathcal{S}$ $\pi(A_t|S_t)$ $\pi(a|s)$ pling ratio $\rho_t = \frac{1}{\mu(A_t|S_t)}$ $\mathbb{E}_{\mu}[\rho_t|S_t = s] = \sum \mu(a|s) \frac{1}{\mu(a|s)} = \sum \pi(a|s) = 1$ $a \qquad \qquad \mu(\omega|\omega)$ and the relationship of taking taking taking taking $\sum_{\mu}(a|s) \frac{\pi(a|s)}{s} Z_{t+1} = \sum_{\pi}(a|s) Z_{t+1} = \mathbb{E}_{\pi}[Z_{t+1}|S_t = s]$ μ _r, σ \sim \sim \sim \sim *µ*(*At|St*) $\mathbf{x}_t \doteq \mathbf{x}(S_t).$ $\left[\left(\gamma \mathbf{x}_{t+1}\right)^\top\right]$ ⇡(*a|s*)=1*.* (8) S $\frac{c}{s}$ case the ratio will be greater or less than one depending on whether $\frac{1}{s}$ key matrix now λ ^T polynical structure $\sum d_u(s) \mathbf{x}(s) \mathbf{x}(s)$ \mathbf{v} and **i** matrices,
it is not stable $\mathbf{x} = \mathbf{X}^T \mathbf{D}$. $(\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X}$. new ergodic stationary distribution $[\mathbf{d}_{\mu}]_s \doteq d_{\mu}(s) \doteq \lim_{t \to \infty} \mathbb{P}\{S_t \!=\! s\} > 0, \forall s \in \mathcal{S}$ states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ we have never the new taken by $\frac{1}{\nu}$ and the sequences we assume that $\frac{1}{\nu}$ and $\frac{1}{\nu}$ target policy $\pi(a|s)$ is no longer used to select actions assume coverage:
behavior policy $\pi(a|s)$ is used to select actional $\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$ old value function $v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t|S_t\!=\!s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ $\pi(A \mid S)$ and behavior points with proportions $\pi(A \mid S)$ importance sampling ratio $\rho_t = \frac{\mu(t) \mathcal{L}_t(\mathcal{S}_t)}{\mu(t) \mathcal{L}_t(\mathcal{S}_t)}$ $\mathbb{E}_{\mu}[\rho_t|S_t = s] = \sum_{\mu} \mu(a|s) \frac{\mu(a|s)}{\mu(s)} = \sum_{\mu} \pi(a|s)$ $\mu(u|s)$ and $\frac{a}{a}$ for $\mu(u|s)$ and $\frac{a}{a}$ For any r.v. Z_{t+1} : $\mathbb{E}_{\mu}[\rho_t Z_{t+1} | S_t = s] = \sum_{\mu} \mu(a|s) \frac{\pi(a|s)}{s} Z_{t+1} = \sum_{\pi} \pi(a|s) Z_{t+1} = \mathbb{E}_{\pi}[Z_{t+1} | S_t = s]$ $+$ ρ_t ϵ $(R_{t+1} +$ *µ*(*At|St*) $\mathbf{x}_t \mathbf{x}_{t+1} - \mathbf{w}_t \mathbf{x}_t \mathbf{x}_t$ $\mathbf{x}_t = \mathbf{x}_t$ $\sum_{\mathbf{b}_t}$ is called the *importance* sampling $\sum_{\mathbf{b}_t}$ and its **A** matrix: $\frac{1}{2}$ $\frac{1}{2}$ ⇡(*a|s*) $\sum d_{\mu}$ (*d* = X \lfloor $\mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top \Big| S_t = s$ $=\sum d_{\mu}(s)\mathbb{E}_{\pi}\big[\mathbf{x}_t\left(\mathbf{x}_t-\gamma\mathbf{x}_{t+1}\right)^{\top}\big|S_t=s\big]$ \mathcal{S} the same steps can be treated the same as in the same as in the on-policy \mathcal{S} case. On other time steps the ratio will be greater $\left\{\begin{array}{ccc} \sim & \sim & \sim & \sim \end{array}\right\}$ has mismatched $\leftarrow \sum d_{\mu}(s) \mathbf{x}(s) \mathbf{x}(s) - \gamma \sum [\mathbf{P}_{\pi}]_{ss'} \mathbf{x}(s')$ **D** and **P** matrices; . $\frac{1}{\sqrt{1-\frac{1}{2}}}$ states $D_t \subset \mathcal{O}$ activities $\Delta t_t \subset \mathcal{J}$ the finite number of $\Delta t_{t+1} \subset \mathbb{R}$ $\begin{array}{ccccccc}\n\text{new encoding stationary distribution} & \text{[d]} & \stackrel{\text{\normalsize{\textbf{d}}}}{ } & \text{[e]} & \stackrel{\text{\normalsize{\textbf{d}}}}{ } & \text{[im.} & \mathbb{P}\{S,-s\} > 0 \end{array} \forall s \in S\n \end{array}$ importance sampling ratio $\rho_t = \frac{\pi (A_t | \beta_t)}{\mu(A + S)}$ $\mathbb{E}_{\mu}[\rho_t | S_t = s] = \sum_{\mu} \mu(s)$ and *Gt*, the *return* at time *t*, is defined by T.V. Z_{t+1} : $\mathbb{E}_{\mu}[\rho_t Z_{t+1} | S_t = s] = \sum_a \mu(a|s) \frac{\pi(a|s)}{\mu(a|s)} Z_{t+1} = \sum_a \pi(a|s) Z_{t+1} = \mathbb{E}_{\pi}[Z_{t+1} | S_t = s]$ linear off-policy TD(0): $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \rho_t \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t$ $\mathbf{x}_t \doteq \mathbf{x}(S_t)$ $\mathbf{w}_t + \alpha \left(\rho_t R_{t+1} \mathbf{x}_t \right)$ $| \sum_{i=1}^n | x_i |^2$ $\left| \cdot \right|$ and its **A** matrix: $\mathbf{A} = \lim_{t \to \infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t \to \infty} \mathbb{E}_{\mu} \left| \rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right) \right|$ The matrix A*^t* multiplies the parameter w*^t* and is thereby critical to the stability of the $\sum_{\mathbf{e}}^{i\alpha}\mu$ (β) $-n$ β -corresponding to corresponding to corresponding of α **I** where $\sum_{s=1}^{\infty} d_{\theta}(s) \mathbf{x}(s) \mathbf{x}(s) = \gamma \nabla$ $\sum_{\mu} \left(\begin{array}{cc} \mu & \mu \\ \mu & \nu \end{array} \right)$ (1) (2) Δ il is fill stable $-\lambda$ $-\mu$ λ μ λ , **DIT-policy Temporal-Difference Learning with Linear Function Approximation**
Les S, E & actions $A_t \in \mathcal{A}$ rewards $B_{t+1} \in \mathbb{R}$ S ^{*j*} $\frac{1}{2}$ vortance sampling ratio $\rho_t \doteq \frac{\mu(A_t|S_t)}{\mu(A_t|S_t)}$ $\qquad \mathbb{E}_{\mu}[\rho_t|S_t\!=\!s] = \sum \mu(a|s) \frac{\mu(a|s)}{\mu(a|s)} = \sum \pi(a|s) = \sum \mu(a|s)$ $\mu(\mathcal{A}t|\mathcal{O}t)$ $=$ \cdot ⇣ α *k* + α (*StR*_{*t*} $\frac{1}{2}$ \mathbf{b}_t $\frac{1}{2} + 1$ **x**_t $\frac{1}{2}$ (**x**_t $\frac{1}{2}$ (**x**_t $\frac{1}{2}$ (**x**_t $\frac{1}{2}$ (**v**_t_{*n*}) ${\bf A}_t$ $\mathbf{A} = \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_{\mu} \left[\rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right) \right]$ $=$ $\overline{ }$ The matrix A*^t* multiplies the parameter w*^t* and is thereby critical to the stability of the $\mathbf{z} = \sum d_{\mu}(s) \mathbb{E}_{\pi} \left[\mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top \middle| S_t = s \right]$ S Key matrix now $\mathbf{I}_{\text{Lip}} = \sum_{d} d_{d}(\mathbf{e}) \mathbf{x}(\mathbf{e}) \left(\mathbf{x}(\mathbf{e}) - \mathbf{e} \sum_{k} \mathbf{P} \mathbf{I}_{d} \mathbf{x}(\mathbf{e}^{\prime}) \right)^{\mathrm{T}}$ has mismatched $\sum u_\mu(s) \mathbf{x}(s) \mathbf{x}(s) - \gamma \sum \mathbf{F}_{\pi}]_{ss'} \mathbf{x}(s')$ **D** and **P** matrices; $\sum_{r=1}^{\infty}$ and $\sum_{r=1}^{\infty}$ are all positive, $\sum_{r=1}^{\infty}$ it is not stable $\mathbf{X}^{\top}(\mathbf{D}_{\mu}(\mathbf{I}-\gamma\mathbf{P}_{\pi})\mathbf{X},$ policy ⇡ : ^A⇥^S ! [0*,* 1], where ⇡(*a|s*) *.* = P*{A^t* =*a|S^t* =*s}*. The state and action sets S and A states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ $v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t|S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ sampling ratio $\rho_t \doteq \frac{\mu(A_t|S_t)}{\mu(A_t|S_t)}$ $\mathbb{E}_{\mu}[\rho_t|S_t = s] = \sum \mu(a|s) \frac{\mu(a|s)}{\mu(a|s)} = \sum \pi(a|s) = 1$ μ (\mathcal{L} **t**_{l}) \mathcal{L} $-w_t + u \left(\underbrace{\rho_t \cdot \mathbf{1}}$ \mathcal{R}_t $\left(\frac{p_t \mathbf{x}_t}{\mathbf{x}_t} \times \frac{p_t \mathbf$ $\frac{t}{\sqrt{2}}$ \mathbb{F} ^[0] ${\bf A}_t$ $\qquad \qquad$ \qquad \qquad and its **A** matrix $\mathbf{A} = \lim_{k \to \infty} \mathbb{E}[\mathbf{A}_k] = \lim_{k \to \infty} \mathbb{E}[\int_{\partial_k} \mathbf{x}_k (\mathbf{x}_k - \gamma \mathbf{x}_{k+1})^\top]$ $\begin{bmatrix} t \to \infty & t' \end{bmatrix}$ (1) and (1) a $=\sum_{\mu} d_{\mu}(s) \mathbb{E}_{\mu}$ $=\sum d_\mu(s) \mathbb{E}_\pi\Big[\mathbf{x}_t\left(\mathbf{x}_t-\gamma\mathbf{x}_{t+1}\right)^\top\Big|S_t=s\Big]$ iteration. To develop intuitivities, consider the special case in which Att is a diagonal matrix. The special matrix \overline{s} $\begin{array}{ccc} \text{key matrix now} & \longrightarrow & \text{if} & \longrightarrow & \end{array}$ has mismatched $=$ $\sum d_{\mu}(s) \mathbf{x}(s) \mathbf{x}(s) - \gamma \sum_i [\mathbf{P}_{\pi}]_{ss'} \mathbf{x}(s')$ **D** and **P** matrices; $\sum_{i=1}^{n}$ $\frac{1}{2}$ importance sampling ratio $\rho_t \doteq$ Off-policy Temporal-Difference Learning with Linear Function App **but-policy** lemporal-Difference Learning with Linear Function App

states S, ⊆ S, actions $A_i \subseteq A$, rewards B_i , $A \subseteq \mathbb{R}$ the actual taken actually encountered, $\begin{bmatrix} \alpha \mu \end{bmatrix} s$ $\alpha \mu \cdot \nu$ and $\alpha \mu \cdot \nu$ at $\alpha \mu \cdot \nu$ $\frac{\pi(A_t|S_t)}{S_t}$ $\mu(A_t|S_t)$ $\mathbb{E}_{\mu}[\rho_t|S_t = s] = \sum_{\mu(a|s)} \frac{\pi(a|s)}{\mu(a|s)}$ For any r.v. Z_{t+1} : $\mathbb{E}_{u}[\rho_{t}Z_{t+1}|S_t = s] = \sum_{u}(a|s) \frac{\pi(u|s)}{s} Z_{t+1} = \sum_{u} \pi(a|s) Z_{t+1} =$ $\mu + \rho$ ^{*i*} (*b*) *µ*(*a|s*) $\mathbf{x}_t\left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t\right) \mathbf{x}_t$ The ratio will be exactly one only on time steps on which the action probabilities for the two case is a matrix. $\mathbf{A} = \min_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \min_{t\to\infty} \mathbb{E}_\mu\left[\frac{pt\mathbf{A}_t}{t} \mathbf{A}_t - \frac{pt\mathbf{A}_t}{t+1} \right]$ $-\sum_{\alpha} d_{\alpha}(\alpha) \mathbb{F} \left[\alpha, \mathbf{x}, (\mathbf{x}_1 - \alpha \mathbf{x}_2, \alpha)^{\top} \right] \mathbf{S}$ ol Difference Leorning with Linear Eunetian Annrevimetian at Binoronoo Eoarning with Enioar Panolion Approximation, *S*
C A $\frac{1}{s}$ $\left\vert s\right\rangle$ \geq \overrightarrow{U} \implies $\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$ *a* $\mu(a|s)$ $\pi(a|s)$ $\mu(a|s)$ $=$ \sum *a* $\pi(a|s)=1$ $\mathbb{E}_{\mu}[\rho_t Z_{t+1} | S_t = s] = \sum_{\mu} \mu(a|s) \frac{\pi(a|s)}{\mu(a|s)} Z_{t+1} = \sum_{\pi} \pi(a|s) Z_{t+1} = \mathbb{E}_{\pi}[Z_{t+1} | S_t = s]$ $\begin{array}{ccc} a & & a \ & & a & \\ & & & \end{array}$ licy TD(0): $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \rho_t \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t$ $\mathbf{x}_t \doteq \mathbf{x}(S_t)$ \mathbf{r} and action taken was more or \mathbf{r} than \mathbf{r} $m_l + \alpha \left(\underbrace{\rho_l \mathbf{r} \mathbf{r}_{l+1} \mathbf{r}_{l}}_{l} \underbrace{\rho_l \mathbf{r}_{l+1} \left(\mathbf{r}_{l} \right) \mathbf{r}_{l+1}}_{l} \right)$ $\pi(A_t|S_t)$ $\pi(a|s)$ sampling ratio $\rho_t = \frac{\partial f(t, y)}{\partial u(x, y)}$ if $\mathbb{E}_{\mu}[\rho_t|S_t = s] = \sum \mu(a|s) \frac{\partial f(t, y)}{\partial u(s)} = \sum \pi(a|s) = 1$ *a* $\mu(a|s)$ $\pi(a|s)$ $\mu(a|s)$ $Z_{t+1} = \sum$ *a* $\mathbb{P}\left[\sup_{t\in\mathcal{A}}\left[\log\left|Z_{t+1}\right|\right]S_{t} = s\right] = \sum_{t\in\mathcal{A}}\mu(a|s)\frac{\mu(a|s)}{\mu(a|s)}Z_{t+1} = \sum_{t\in\mathcal{A}}\pi(a|s)Z_{t+1} = \mathbb{E}_{\pi}[Z_{t+1}|S_{t} = s].$ $\mathbf{S} = \mathbf{W}_t + \alpha \begin{pmatrix} 0 & R_{t+1} \mathbf{Y}_t - \alpha \mathbf{Y}_t (\mathbf{Y}_t - \alpha \mathbf{Y}_{t+1})^\top \mathbf{W}_t \end{pmatrix}$ \mathbf{b}_t **A**_t

lim $\mathbb{F}[\mathbf{A}_t] = \lim_{k \to \infty} \mathbb{F} \left[\arg \left(\mathbf{x}_t + \cos \theta_t \right)^\top \right]$ $\frac{1}{2}$ \overline{z} $d_\mu (s) \mathbb{E}_\mu \, | \,$ b*t* \int_{0}^{t} \mathbf{x}_{t} $(\mathbf{x}_{t} - \gamma \mathbf{x}_{t+1})^{\top}$ $\vert S \vert$ A*t* w*^t ,* is never taken under the target policy in the target policy in that state on update on \mathcal{S} that step matrix now this algorithm \sim has mismatched \le \ge \ge not stab it is not stable \overline{a} $\mathbf{X}^{\mathsf{T}}\left(\mathbf{D}_{\mu}(\mathbf{I}-\gamma\mathbf{P}_{\pi})\mathbf{X},\right)$ nce Learning with Linear Function Approximation bolicy $\mu(u|s)$ is used to select actions!
 $\pi(u|s) > 0 \implies \mu(u|s) > 0 \implies \pi(u|s) > 0$ godic stationary distribution $[\mathbf{d}_\mu]_s\doteq d_\mu(s)\doteq \lim_{t\to\infty}\mathbb{P}\{S_t\!=\!s\}>0, \forall s\in\mathcal{S}$ π ⁽³) $\frac{1}{2}$ $|f(S_t)|$ *µ*(*a|s*) *z*^{*t*}+1 $\frac{1}{2}$ *a* $\pi(a|s)$ We can use this fact to begin to adapt TD(0) for o↵-policy learning (Precup, Sutton & $S \cap V$. Z_{t+1} : $\mathbb{E}_{\mu}[\rho_t Z_{t+1} | S_t = s] = \sum_{\mu} \mu(a|s) \frac{\sum_{t=1}^s Z_{t+1}}{\mu(a|s)} Z_{t+1} = \sum_{t=1}^s \pi(a|s) Z_t$ $\sqrt{2}$ $R_{t+1} + \gamma \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$ blicy TD(0): $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \rho_t \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t$ $\qquad \mathbf{x}_t \doteq \mathbf{x}(S_t)$ $=$ **w**_t + α $\sqrt{2}$ $\rho_t R_{t+1} \mathbf{x}_t$ $\bigcup_{\mathbf{b}}$ \mathbf{b}_t $-\rho_t\mathbf{x}_t\left(\mathbf{x}_t-\gamma\mathbf{x}_{t+1}\right)^\top$ \overbrace{A} . \mathbf{A}_t w*^t* $\sum_{i=1}^{n} \mathbf{r}_i = \mathbf{r}_i \mathbf{r}_i + \mathbf{r}_i \mathbf{r}_i$ and its \mathbf{A} matrix: $\mathbf{A} = \lim_{t \to \infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t \to \infty} \mathbb{E}_{\mu} \left[\rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^{\top} \right]$ $\mathbf{v} = \sum d_{\mu}(s) \mathbb{E}_{\mu} \left[\rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top \right] S_t = s$ that step, as desired. We call this algorithm *o*↵*-policy TD(0)*. k av m **t** $\left\{\n \begin{array}{ccc}\n \cdot & \cdot & \cdot \\
 \hline\n \end{array}\n\right\}$ \mathbf{S} $ed \leq \sum a_{\mu}$ \int $\mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \sum_{s}$ $\frac{1}{2}$ $\mathcal{S}_{\mathcal{S}}$ *s*
S
S s. *dµ*(*s*)E⇡ x*^t* (x*^t* x*t*+1) $x^2 + y^2 = 0$ assume coverage ⇡(*a|s*) $\sigma_t^{\top} \mathbf{x}(s)$ $\sum_{\mu(a|s)} \frac{\pi(a|s)}{s}$ $\sum \pi(a|s) = 1$ = w*^t* + ↵ ⇢*tRt*+1x*^t* $\frac{1}{2}$ $\mathbb{E}_{\pi}[Z_{t+1}]$ b*t* ⇢*t*x*^t* (x*^t* x*t*+1) $=$ S v_{t+1} **x**_t $-\rho_t$ **x**_t (**x**_t $-\gamma$ **x**_{t+1}) **w**_t A_t $\frac{1}{2}$ $-\gamma \mathbf{x}_{t+1}$ ^T $S_t = s$ []] $\frac{1}{2}$ *dµ*(*s*)E*^µ* \sim *dµ*(*s*)E⇡ = X *dµ*(*s*) x(*s*) $A_1 \in \mathbb{R}$ $\pi(a|a)$ $_{t}|S_{t})$ $\sum_{a}^{\mu} \mu[t|bt|bt-s] = \sum_{a}^{\mu} \mu(a|s) \mu(a|s)$ ${\pi(a|s)}$ \boldsymbol{a} \overline{a} $\overline{1}$ ⌘ *,* where $\frac{a}{a}$, $\mu(a|s)$ $\begin{array}{ccc} a & & a \end{array}$ that step, as desired. We call this algorithm *o*↵*-policy TD(0)*. $= \mathbf{W}_t + \alpha \left(\rho_t R_t \right)$ ${\bf A} = \lim$ $\lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_{\mu}$ $\sqrt{ }$ $\rho_t \mathbf{x}_t\left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)$ T $=$ \sum *s* $d_\mu(s)\mathbb{E}_\mu$ $\sqrt{ }$ $\rho_t \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top$ $\overline{}$ \vert $S_t = s$ $\overline{}$ $=\sum d_{\mu}(s)\mathbb{E}_{\pi}$ *s* $\sqrt{ }$ $\mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1} \right)^\top \Big|$ $\begin{array}{c} \end{array}$ $S_t = s$ $\overline{1}$ $=$ \sum *s* $d_{\bm{\mu}}(s) \, \mathbf{x}(s)$ $\sqrt{ }$ $\mathbf{x}(s) - \gamma$ \sum *s*0 $[\mathbf{P}_\pi]_{ss'} \mathbf{x}(s')$ Δ key matrix now has mismatched

from previous definitions. In on-policy training, the actions are chosen according to a target

parameter vector ^w*^t* ² ^R*n*, made at each of a sequence of time steps *^t* = 0*,* ¹*,* ²*,...*, on

In general, for any random variable *Zt*+1 dependent on *St*, *A^t* and *St*+1, we can recover

parameter 2 [0*,* 1). Conventional linear TD(0) is defined by the following update to the

parameter 2 [0*,* 1). Conventional linear TD(0) is defined by the following update to the

sums to <0! \blacksquare \blacksquare \blacksquare *Iference Learning with Linear Func* states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ *s dononel*
 dions! target policy $\pi(a|s)$ is no longer used to select actions assume coverage: $\pi(a|s) > 0$ $\pi(a|s) > 0 \implies \mu(a|s) > 0$ $\forall s,a$ $\alpha \textbf{w}_t^\top \textbf{x}(s)$ *s* $1t\rightarrow\infty$ \perp \vee t \rightarrow \circ f \lt \cup , \vee \circ \lt \circ X^{\pm} s) key matrix now $\mathcal{L}(w)$ $\left(\frac{2w}{\pi(\text{right}|\cdot)}\right)$ $\frac{\mu(\text{right}|\cdot)}{\pi(\text{right}|\cdot)}$ = 1 where D*^µ* is the *N* ⇥ *N* diagonal matrix with the stationary distribution d*^µ* on its diagonal. Thus, the key matrix that must be positive definite is D*µ*(I P⇡) and, unlike in the onpolicy case, the distribution and the transition probabilities do not match. We do not have $\gamma = 0.9$ $\bigcup_{\alpha} \bigcup_{\alpha} \bigcup_{\alpha} \bigcup_{\alpha} \bigcup_{\alpha} \bigcup_{\alpha} \bigcup_{\alpha} \bigcap_{\alpha} \bigcap$ A simple *w*!2*w* example of divergence that fits the setting in this section is shown in prob matrix: $P_{\pi} = \begin{bmatrix} 0 & 1 \end{bmatrix}$ $|P_{\pi}|_{ij} = \sum_{a} \pi(a|i)p(j|i,a)$ left or right states. All the rewards are zero. As before, there is a single parameter *w* and $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $\mathbf{D}_{\mu}(\mathbf{I}-\gamma\mathbf{P}_{\pi})=\begin{bmatrix}0.0 & 0\\ 0 & 0\end{bmatrix}\times\begin{bmatrix}1 & -0.9\\ 0 & 0.1\end{bmatrix}=\begin{bmatrix}0.0 & -0.40\\ 0.05\end{bmatrix}$ sums to <0! states, such that equal time is spent on average in both states, d*^µ* = (0*.*5*,* 0*.*5)>. The target policy is to go right in both states. We seek to learn the value from each state given that estion is $\mathbf{Y}^\top \mathbf{D}$ if an example is continued. The this example is the this example is the this example is $|0.5$ and $-0.45|$ and $|1|$ and $|1|$ and $|1|$ and $|0.4|$ and $|0.4|$ and $|1.4|$ and $|0.4|$ and $|0.$ b^{ll} parameter 2 [0*,* 1). Conventional linear TD(0) is defined by the following update to the **Off-policy Temporal-Difference Learning with Linear Function Approximation** $P^{\text{tr}(\alpha|S)}$ is used to select actions:
bw ergodic stationary distribution $[\mathbf{d}_{\mu}]_s \doteq d_{\mu}(s) \doteq \mathrm{Id}_s$ $\mu(s)$ is used new ergodic stationary distribution $[\mathbf{d}_{\mu}]_s \doteq d_{\mu}(s) \doteq \lim_{t \to \infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in \mathcal{S}$ ω value function $v_{\pi}(s) = \mathbb{E}_{\pi}[\mathbf{G}_t | \beta_t = s] \approx \mathbf{w}_t \mathbf{x}(s)$ key matrix now state *^s*. The notation " *.* =" indicates an equality by definition rather than one that follows =" indicates an equality by definition rather than one that follows $f = 0.9$ $T(\text{right}|\cdot)$ $\begin{matrix}\n \text{Counterexample:} & \lambda = 0 & \quad \text{if} & \text{if}$ \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc 5 are assumed to be finite, but the number of $\sum_{i=1}^N \sum_{j=1}^N \sum_{j=$ $\gamma = 0.9$ $\left(\begin{array}{c} w \ w \end{array}\right)$ $\left(\begin{array}{c} 2w \ w \end{array}\right)$ $\left(\begin{array}{c} \mu_{\text{(right[-]})} - 0.5 \text{ |g||l|}^2) - \mathbf{X} = 0.9$ linear function approximation, in which the inner product of the parameter vector and the $\begin{bmatrix} 0 & 1 \end{bmatrix}$ transition prob m
 x SILION DIGITS $\mathbf{r} = \begin{bmatrix} 0 & 1 \end{bmatrix}$ $\mathbf{r} = \begin{bmatrix} 1 & \pi \end{bmatrix}$ $\mathbf{r} = \begin{bmatrix} 1 & \pi \end{bmatrix}$ key matrix: $\mathbf{D}_{\mu}(\mathbf{I} - \gamma \mathbf{P}_{\pi}) = \begin{vmatrix} 0.5 & 0 \ 0 & 0.5 \end{vmatrix} \times \begin{vmatrix} 1 & -0.9 \ 0 & 0.1 \end{vmatrix} = \begin{vmatrix} 0.5 & -0.45 \ 0 & 0.05 \end{vmatrix}$ sum $\mathbf{A}\mathbf{B}$: $\mathbf{X}^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} -1 & 0 & 0 \end{bmatrix}$ parameter 2 [0*,* 1). Conventional linear TD(0) is defined by the following update to the from previous definitions. In our chosen actions are chosen actions a α be another number of α of α is assumed much in the number of α $\lambda = 0$
 $\gamma = 0.9$ $\left(\sqrt{w} \sum_{n=0}^{\infty} w \right)^n$ λ linear function approximation, in the parameter vector \sim Γ eature vector for a state is meant to that state is meant to the value of that state Γ \bm{p} pos def test: $\mathbf{X}^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & -0.45 \ 0 & 0.25 \end{bmatrix} \times \begin{bmatrix} 1 \ 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} -0.4 \ 0.1 \end{bmatrix} = -0.2$ parameter vector ^w*^t* ² ^R*n*, made at each of a sequence of time steps *^t* = 0*,* ¹*,* ²*,...*, on $u_{\mu}(s) = \min_{t \to \infty} \mathbb{E}\{\partial_t = s\} > 0, \forall s \in \mathcal{S}$ old value function $v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t|S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ $\lambda = 0$ \sim \sim \sim \sim μ (right $|\cdot$) = 0.5 $)= 1$ are assumed to be finite the number of states is assumed much larger than the number of states is assumed much larger than the number of states is assumed much larger than the number of states is assumed much larger than t $\bigcap_{\Omega \in \mathcal{N}} \bigcap_{\mu(\mathsf{rig})}$ $\mu(\text{right}|\cdot) = 0.5$ $\mu(r|\phi| + \mu(r|\phi| + 1) = 0.5$ joht \vdots of the parameter vector $\mu(r|\phi| + 1) = 0.5$ joht \vdots of the parameter vector and the $f(x) = \sum_{\pi(\text{right}|\cdot)} \mu(\text{right}|\cdot) = 0.5$ ight $|\cdot) = \frac{1}{\mathbf{X}} = \begin{bmatrix} 1 \end{bmatrix}$ $[0 \quad 1]$ $[$ $[$ $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $[$ $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ $[$ $\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}]$ $[$ $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ $[$ $\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}]$ $[$ $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}]$ $[$ and *Gt*, the *return* at time *t*, is defined by $\begin{bmatrix} 0.05 \end{bmatrix}$ sums to <0! $in point$ is important binoronoo Lourining with</u> states $D_t \in \mathcal{O}$ actions $\Lambda_t \subset \Lambda$ rewards $n_{t+1} \subset \mathbb{R}$ behavior policy $\mu(a|s)$ is used to select actions! $\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$ $\text{new ergodic stationary distribution } [d_u]_s = d_u(s) = \lim_{t \to \infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in S$ ferent the target policy. $\lambda = 0$ but pointly the relativity of $\gamma = 0.9$ taking the action actually taken, *At*, in the state actually encountered, *St*, under the target $\lambda = 0$ $\left(\frac{1}{2}I\right)^{1/2}$ ⇡(*a|s*) = X ⇡(*a|s*)=1*.* (8) \mathcal{L}_{μ} one only one only one only \mathcal{L}_{μ} (case. On other time steps the ratio $\sqrt{0.5}$ pos del test. $\mathbf{A} \mathbf{D}_{\mu} (\mathbf{I} - \gamma \mathbf{F}_{\pi}) \mathbf{A} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 1 \end{bmatrix}$ $\mathcal{A}_t \in \mathcal{A}$ in rewards $R_{t+1} \in \mathbb{R}$ $i \in \mathcal{U}$ is ignored the values that $i \in \mathcal{U}$ if the action state $i \in \mathcal{U}$ \mathcal{A}_μ (*s*) $\dot{=}$ $\lim_{t\to\infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in \mathcal{S}$ $f(x)$ for example, some action by $\{2w\}$ might never by $\{2w\}$ might never be chosen by $\{2w\}$ might never be chosen by $\{2w\}$ by ⇡. To address this, we use importance sampling to correct for the relative probability of taking the action actually taken, *At*, in the state actually encountered, *St*, under the target $\alpha = 0$ π (right $|\cdot$ $\vec{X} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$
 $\pi(\text{right} \mid \cdot) = 1$ $\vec{X} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ This quantity is called the *importance sampling ratio* at time *t*. Note that its expected value is motrived in the interest. *a a µ*(*a|s*) $\mathbf{D}_{\mu}(\mathbf{I} = \gamma \mathbf{I} \pi) = \begin{bmatrix} 0 & 0.5 \end{bmatrix}$ \times $\begin{bmatrix} 0 & 0.1 \end{bmatrix}$ $\begin{bmatrix} - & 0 & 0.05 \end{bmatrix}$ suris to \leq 0. policies are exactly the same; these time steps can be treated the same as in the on-policy $\begin{bmatrix} 0 & \epsilon & 0 & 4\epsilon \end{bmatrix}$ $\begin{bmatrix} 1 & 1 & \epsilon & 0 & 4 \end{bmatrix}$ $\mathbf{X}^{\top}\mathbf{D}_{\mu}(\mathbf{I}-\gamma\mathbf{P}_{\pi})\mathbf{X} = \begin{bmatrix}1 & 2\end{bmatrix}\times\begin{bmatrix}0.9 & 0.79\\ 0 & 0.05\end{bmatrix}\times\begin{bmatrix}1\\ 2\end{bmatrix} = \begin{bmatrix}1 & 2\end{bmatrix}\times\begin{bmatrix}0.7\\ 0.1\end{bmatrix} = -0.2$ policy, and some kind of correction is neglected. The correction is neglected in Γ or Γ new ergodic stationary distribution $[\mathbf{d}_{\mu}]_s \doteq d_{\mu}(s) \doteq \lim_{t \to \infty} \mathbb{P}\{S_t \!=\! s\} > 0, \forall s \in \mathcal{S}$ states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ we have never the new taken by $\frac{1}{\nu}$ and the sequences we assume that $\frac{1}{\nu}$ and $\frac{1}{\nu}$ target policy $\pi(a|s)$ is no longer used to select actions assume coverage:
behavior policy $\pi(a|s)$ is used to select actional $\pi(a|s) > 0 \implies \mu(a|s) > 0 \quad \forall s, a$ old value function $v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t|S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ $E = 0$ and $E = 0$ and $E = 0$ is covered to be active policy with proportions difoff-policy TD(0)'s $A^{\lambda} = 0$
and $\lambda = 0$ and $\lambda = 0$ and $\lambda = 0$ and $\lambda = 0$. $\gamma = 0.9$ $\sqrt{2}$ $\sqrt{2}$ π (right $\gamma = 1$ taking the action actually taken, *At*, in the state actually encountered, *St*, under the target bournoroxample. ⇢*^t .* \overline{a} $\left[2w\right]$ *µ*(*At|St*) $\begin{bmatrix} 0 & 1 \end{bmatrix}$ is called the *in* time $\begin{bmatrix} 1 & 1 \end{bmatrix}$ transitio *a* \int \int \int \int \int *µ*(*a|s*) $\frac{1}{2}$ ⇡(*a|s*)=1*.* (8) The ratio will be exactly one only on time steps on which the action probabilities for the two policies are exactly the step the same step the step $\begin{bmatrix} 0 & 5 & -0.45 \end{bmatrix}$ [1], $\begin{bmatrix} -0.4 \end{bmatrix}$ pos det test: $\mathbf{X}^{\top} \mathbf{D}_{\mu} (\mathbf{I} - \gamma \mathbf{P}_{\pi}) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 0.05 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 1 \end{bmatrix} =$ the action taken was more or less likely under the target policy than under the behavior Λ is needed. Some kind of correction is needed. . $\frac{1}{\sqrt{1-\frac{1}{2}}}$ but to but the number of \mathcal{L}_t of \mathcal{L}_t is assumed multiple number of \mathcal{L}_t new ergodic stationary distribution $[d_{\nu}]_s \doteq d_{\nu}(s) \doteq \lim_{t \to \infty} \mathbb{P}\{S_t = s\} > 0, \forall s \in S$ $\frac{1}{2}$, the $\frac{1}{2}$ contrary $\frac{1}{2}$ $\begin{array}{ccc} \text{C} \text{counterexample:} & \lambda = 0 & \text{if } \mathcal{A} \leq 0 \end{array}$ $\lambda = 0$
 $\gamma = 0.9$ $\alpha = 0.2$ $\alpha = 0.5$ ight $|\cdot$) = $\frac{1}{X} = \begin{pmatrix} 1 & 1 \ 0 & 1 \end{pmatrix}$ b*t*2R*ⁿ* $\lceil 0 \rceil$ **a** $\begin{bmatrix} 0 & 1 \end{bmatrix}$ $\overline{}$ $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ are $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ are $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ are $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ are $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ are $\text{tr} \tan \theta$ and $\text{tr} \tan \theta$ a Rey matrix: $D_{\mu}(1-\gamma P_{\pi}) = \begin{vmatrix} 0 & 0 \\ 0 & 0 \end{vmatrix} \times \begin{vmatrix} 0 & 0 \\ 0 & 1 \end{vmatrix} = \begin{vmatrix} 0 & 0 \\ 0 & 1 \end{vmatrix}$ iteration. To develop intuition, consider the special case in which A*^t* is a diagonal matrix. Γ and diagonal elements are negative, then the corresponding diagonal elements are negative, then the corresponding diagonal element of Γ pos def test: $\mathbf{X}^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ will lead to divergence if continued. The second term (⊥^btinued. The second term (2¹ *DII-policy* lemporal-Difference L
Les S, E & actions $A_t \in A$ reward $P_{t+1} \in \mathbb{R}$
 $R_{t+1} \in \mathbb{R}$ $a\circ \omega_t \subset \sigma$ activites ι_t of the number of ι_{t+1} on ω off-policy TD(0)'s $A \sim 0.0$ and $\left(\begin{array}{c} \Delta w \\ w \end{array} \right)$ and $\left(2w \right)$ and $\left(\begin{array}{c} \mu_{\text{rel}} \\ \mu_{\text{rel}} \end{array} \right)$ and $\gamma = 0.9$ *^G^t .* $\begin{bmatrix} 0 & 1 \end{bmatrix}$ $\left| \begin{array}{c} 1 \\ 1 \end{array} \right|$ $\sqrt{2}$ $\sqrt{2}$ $\sqrt{2}$ $\sqrt{2}$ $\sqrt{2}$ $\sqrt{2}$ $\frac{1}{2}$ \sum *n* $\begin{bmatrix} 0 & 1 \end{bmatrix}$ (bttp://www.fractional.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/stational.org/statio $\sqrt{10.5}$ $\begin{bmatrix} \mu & 0 \\ \mu & 0 \end{bmatrix}$ and is the parameter with $\begin{bmatrix} 0 & 0.5 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0.1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0.05 \end{bmatrix}$ \mathbf{v} and \mathbf{v} of \mathbf{v} \mathbf{v} of \mathbf{v} or \mathbf{v} are negative, then the corresponding diagonal element of \mathbf{v} \mathbf{v} of \mathbf{v} \mathbf{v} of \mathbf{v} \mathbf{v} of \mathbf{v} \mathbf{v} \mathbf{v} of \mathbf \int ² definest. \mathbf{A} \mathbf{D}_{μ} ($\mathbf{I} = \int \mathbf{A} \pi / \mathbf{A} = \begin{bmatrix} 1 & 2 \end{bmatrix}$ \wedge $\begin{bmatrix} 0 & 0.05 \end{bmatrix}$ \wedge $\begin{bmatrix} 2 \end{bmatrix}$ $\begin{bmatrix} 1 & 2 \end{bmatrix}$ \wedge $\begin{bmatrix} 0.1 \end{bmatrix}$ \mathbf{I} \mathbf{I} which will lead to divergence if continued. The second term (1⊥bt) does not a
Does not a<mark>→ect the second term (</mark>1⊥bt) does not a</u> from previous definitions. In on-policy training, the actions are chosen according to a target Off-policy Temporal-Difference Learning with Linear Function Approximation states $S_t \in \mathcal{S}$ actions $A_t \in \mathcal{A}$ rewards $R_{t+1} \in \mathbb{R}$ old value function $v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t|S_t = s] \approx \mathbf{w}_t^{\top} \mathbf{x}(s)$ key matrix now cy TD(0)'s $A \frac{\lambda - \upsilon}{\alpha - \alpha}$ ($\chi \gamma$) Qw) $A^{(1)}$ (right) γ $\gamma - 0.9$ \sum w*t*+1 = w*^t* + ↵ *Rt*+1x(*St*) $[{\bf P}_x]_{ii}$ $\left[{\bf P}_\pi\right]$ $\sum \pi(a|i) n(i|i, a)$ $\mathbf{P}_{\pi} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ $[\mathbf{P}_{\pi}]_{ij} \doteq \sum_{a} \pi(a|i)p(j|i,a)$ = w*^t* + ↵(b*^t* A*t*w*t*) (4) $(t) = \begin{vmatrix} 0.5 & 0 \\ 0 & 0 \end{vmatrix} \times \begin{vmatrix} 1 \\ 0 \end{vmatrix}$ The matrix A*^t* multiplies the parameter w*^t* and is thereby critical to the stability of the τ is the special constant $\begin{bmatrix} 0.5 & -0.45 \end{bmatrix}$ [1], a diagonal matrix. SI: $\mathbf{X} \cdot \mathbf{D}_{\mu} (\mathbf{I} - \gamma \mathbf{P}_{\pi}) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 0.05 \end{bmatrix} \times \begin{bmatrix} 0.05 \end{bmatrix} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 1 \end{bmatrix} = -0.2$ I↵A*^t* will be greater than one, and the corresponding component of w*^t* will be amplified, which we estimate the second term (10 to divergence term (10 to does not appel) off-policy TD(0)'s $A_{\alpha = 0.0}^{\lambda = 0}$ (χ and χ and χ and μ (right) \mathcal{D} and π (right). \mathcal{D} it is not stable $\left($ Figure 1. From each state there are two actions, which take the process to the process the single feature is 1 and 2 in the two states such that the approximate values are *w* and sources, such that equal time is spent on a vertex, γ_{11} γ_{21} γ_{31} , γ_{11} , γ_{21} , γ_{11} , γ_{21} , γ_{11} , γ_{21} , γ_{21} , γ_{11} , γ_{21} , γ_{21} , γ_{21} , γ_{21} , γ_{21} , γ_{21} , $\lambda = 0$ $\lambda = 0$ $\gamma = 0.9$ $\sqrt{2}$ $\pi(rvert;\cdot) = 1$ $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ α abuting coverage.
 α and α approach to α and α α $\left\langle w \right\rangle$ $\left\langle 2w \right\rangle$ π (right) = 1 the state and initial transient, states with states with the state of the state of the state of the state distribution of the state μ (right) = 0.5
 μ (right) = 0.5
 μ (right) = 0.5 $w = \left(2w - \frac{1}{\pi(\text{right})}-1\right)$ is positive integrated with $\bar{\mathbf{X}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ $\begin{bmatrix} 1 & 2 \end{bmatrix}$. The steady- $\begin{bmatrix} 0 & 1 \end{bmatrix}$ $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ $[{\bf P}_{\pi}]_{ij} \doteq \sum_a \pi(a|i)p(j|i,a)$ $\begin{bmatrix} 0 & -1 \end{bmatrix}$ 5] $\begin{bmatrix} 0 & 0.5 \end{bmatrix}$ $\begin{bmatrix} 0 & 0.1 \end{bmatrix}$ $\begin{bmatrix} 0 & 0.05 \end{bmatrix}$ $\begin{bmatrix} 1 & 2 \end{bmatrix}$, $\begin{bmatrix} -0.4 \end{bmatrix}$ 0.9 $\mathbf{A} = \begin{bmatrix} 1 & 2 \end{bmatrix}$ \wedge $\begin{bmatrix} 0 & 0.05 \end{bmatrix}$ \wedge $\begin{bmatrix} 2 \end{bmatrix}$ = lim*t*!1 E[b*t*]. We say that the algorithm and its expected update are *stable* i↵ the A matrix is positive definite, and that they are *unstable* if A is transition prob matrix: $\mathbf{A}(s) = \mathbb{E}_{\pi}[\mathbf{G}_t | S_t \!=\! s] \approx \mathbf{W}_t^{\top} \mathbf{X}(s)$ key matring $\lambda = 0$ $\lambda = 0$ $\bigcap_{u \in \mathcal{U}} \bigcap_{v \in \mathcal{V}} \bigcap_{v \in \mathcal{V}} \bigcap_{v \in \mathcal{V}} \bigcap_{v \in \mathcal{V}} \mu(\text{right}|\cdot) = 0.5$ $\lambda = 0$
 $\lambda = 0$ $\mu(\text{right}|\cdot) = 0.5$ $\mu(\text{right}|\cdot) = 0$ Figure 1: *w*!2*w* example without a terminal state. The key matrix is $\mathbf{D}_{\mu}(\mathbf{I} - \gamma \mathbf{P}_{\pi}) = \begin{bmatrix} 0.5 & 0 \ 0 & 0.5 \end{bmatrix}$ 0 0*.*5 $\overline{}$ ⇥ $\begin{bmatrix} 1 & -0.9 \end{bmatrix}$ 0 0*.*1 $\overline{}$ = $\begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix}$ sums to $<$ 0! $\begin{bmatrix} 0 & 5 & -0.45 \end{bmatrix}$ [1] $\begin{bmatrix} 0 & 4 \end{bmatrix}$ $\text{dist} \text{ is a set: } \quad \mathbf{X}^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & -0.45 \ 0 & 0.05 \end{bmatrix} \times \begin{bmatrix} 1 \ 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} -0.4 \ 0.1 \end{bmatrix} = -0.2$ it is not positive definite by multiplying it on both sides by y = X = (1*,* 2)>: key matrix: pos def test: And to object the continuous $\pi(a|a) \times 0 \longrightarrow \mu(a|a)$ $\gamma = 0.9$ $\left(\sqrt{w}\right)$ $\lambda = 0$ $\bigwedge_{uv} \bigwedge_{v} \bigwedge_{v} \bigwedge_{v} \bigwedge_{v} \mu(\text{right}|\cdot) = 0.5$ w $\left(\frac{2w}{\pi(right|\cdot)}\right)$ $\frac{\mu(right|\cdot)}{\pi(right|\cdot)}$ = 1 $\gamma = 0.9$ $\mathbf{P}_{\pi} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ $[\mathbf{P}_{\pi}]$ $\lfloor \mathbf{P}_\pi \rfloor_i$ $\frac{1}{\sqrt{1-\frac{1}{2}}}$ 1 0*.*9 $\sum_a \pi(a)$ \overline{n} ix: $\mathbf{P}_{\pi} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad [\mathbf{P}_{\pi}]_{ij} \doteq \sum_{a} \pi(a|i)p(j|i,a)$ $\begin{bmatrix} 0 & r & 0 & 1 \end{bmatrix}$ is the contraction to $r = 0$ matrix matrix matrix matrix matrix matrix matrix matrix matrix $\begin{bmatrix} 0 & r & 0 & 0 \end{bmatrix}$ that the second column sums to a negative number $D_{\mu}(I - \gamma P_{\pi}) = \begin{vmatrix} 0.0 & 0 \\ 0 & 0.5 \end{vmatrix} \times \begin{vmatrix} 1 & -0.9 \\ 0 & 0.1 \end{vmatrix} = \begin{vmatrix} 0.0 & -0.40 \\ 0 & 0.05 \end{vmatrix}$ sums to <0! it is not positive definite by multiplying it on both sides by y = X = (1*,* 2)>: $\begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix} \times$ $\lceil 1 \rceil$ 2 $\overline{}$ $= \begin{bmatrix} 1 & 2 \end{bmatrix} \times$ $\sqrt{-0.4}$ 0*.*1 $\overline{}$ $=-0.2$ \blacksquare That this is negative means that the key matrix is not positive definite. We have also be have also been w $(2w)$ π (right|·) = 1 π $\mathbf{D}_{\mu}(\mathbf{I} - \gamma \mathbf{P}_{\pi}) = \begin{bmatrix} 0 \end{bmatrix}$ $= \begin{array}{c} \hline \hline \end{array}$ $\overline{5}$ 0¹. $5^{\frac{\times}{2}}$ $\overline{\Gamma}$ = $\begin{bmatrix} -0.9 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix}$.₀ | ^{-0.40} | sum that the second sums to a negative number of $[0.5$ $-0.45]$ $[1]$ that a show that s definest. $\mathbf{X} \cdot \mathbf{D}_{\mu} (\mathbf{I} - \gamma \mathbf{F}_{\pi}) \mathbf{X} = \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \end{bmatrix} = 0$ **A** → ⇒ ية المسيحة العامة ا
العامة العامة العا Counterexample: $\lceil 1$ 2 $\overline{}$

⌘ .
y A is not positive definite! Stability is not assured. all diagonal elements between 0 and 1. In this case the first term of the update tends to 11 Person smaller than the largest of the than the largest of the than the largest of the I is diagonal with the internal with the such that I is discussed to the internal with the such that I is discussed to the internal A is not positive definite. Stability is not assured. definite! Stability is $\overline{101}$ dood \mathbf{r} A is not positive definite: Stability is not assured. $\mathbf A$ 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 $GTD(\lambda)$ and $GQ(\lambda)$

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…

Ben Van Roy 2009

Other Distribution

Ben Van Roy 2009

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 futurereweighted 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 $TD(\lambda)$

Emphatic TD(0) Emphatic TD(0) *s*

Introduces a new short-term memory random variable—the *followon trace:* t is the variance will be much less than in the simplest case of t is the simplest case of t an out es a new snort-temphory random variable—the *followon trace:* proportional to a new short-term memory random variable *F*
F¹ duces a new short-term memory random variable—the *followon trace:* i ahle $-$ *the followon trace:* $F_i \doteq \gamma \rho_{i-1} F_{i-1} + 1$ and $F_{i-1} = 0$ and $F_{i-1} = \gamma$ and $F_{i-1} = 0$ $Intoc$ m variable—the *followon trace:*

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

Emphatic TD(0):

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

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

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

$$
\text{where } \mathbf{F} = \begin{bmatrix} \mathbf{v} & \mathbf{0} \\ \mathbf{0} & \mathbf{v} \end{bmatrix} \qquad \sum_{i} [\mathbf{F} (\mathbf{I} - \gamma \mathbf{P}_{\pi})]_{ij} = \sum_{i} \sum_{k} [\mathbf{F}]_{ik} [\mathbf{I} - \gamma \mathbf{P}_{\pi}]_{kj}
$$

$$
\text{with } [\mathbf{f}]_s = d_{\mu}(s) \lim_{t \to \infty} \mathbb{E}_{\mu} [F_t | S_t = s] \qquad \text{column of } j\text{th}
$$

$$
= \sum_{i} [\mathbf{F}]_{ii} [\mathbf{I} - \gamma \mathbf{P}_{\pi}]_{ij}
$$

$$
= \sum_{i} [\mathbf{f}]_{i} [\mathbf{I} - \gamma \mathbf{P}_{\pi}]_{ij}
$$

$$
= [\mathbf{d}_{\mu}^{\top} (\mathbf{I} - \gamma \mathbf{P}_{\pi})]_{j}
$$

$$
= [\mathbf{d}_{\mu}^{\top} (\mathbf{I} - \
$$

^F^t . = ⇢*t*1*Ft*¹ + 1*,* 8*t >* 0*,* (12) Emphatic TD(0) F *mnhatic*; $TD(0)$ R *r* σ $\mathbf{D}(\mathbf{0})$ Γ _{pectation} $\Gamma \cap (\cap)$ **E***mphatic* I $D(0)$

Introduces a new short-term memory random variable—the *followon trace:* t is the variance will be much less than in the simplest case of t is the simplest case of t an out es a new snort-temphory random variable—the *followon trace:* proportional to a new short-term memory random variable *F*
F¹ *f*(*iable*—the \mathcal{L} uces a new short-term memory random variable—the *followon trace: S^t* = *s* Introduces a r Υ **n** memory randon *S^t* = *s* hort-term memory random varia $\ddot{}$ memory random variable—the *fo* $\overline{11}$ = X>D*µ*(I P⇡)X*,* (10)

Introduces at new short-terminmentory rationality and the-time nonowor trace.

\n
$$
F_{t} \doteq \gamma \rho_{t-1} F_{t-1} + 1, \quad \forall t > 0 \qquad F_{-1} = 0
$$
\nExponential function for $V_{t+1} = 0$

\n
$$
\mathbf{A} = \lim_{t \to \infty} \mathbb{E}[\mathbf{A}_{t}] = \lim_{t \to \infty} \mathbb{E}_{\mu} \Big[F_{t} \rho_{t} \mathbf{x}_{t} \left(\mathbf{x}_{t} - \gamma \mathbf{x}_{t+1} \right)^{\top} \Big] = \mathbf{X}^{\top} \mathbf{F} (\mathbf{I} - \gamma \mathbf{P}_{\pi}) \mathbf{X}
$$
\nwhere $\mathbf{F} \doteq \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ $\begin{array}{c} \text{Counterexample:} \\ \lambda = 0 \\ \gamma = 0.9 \end{array}$

\nwith $[\mathbf{f}]_{s} \doteq d_{\mu}(s) \lim_{t \to \infty} \mathbb{E}_{\mu} [F_{t} | S_{t} = s]$

\nwith $[\mathbf{f}]_{s} \doteq d_{\mu}(s) \lim_{t \to \infty} \mathbb{E}_{\mu} [F_{t} | S_{t} = s]$

\nUsing above:

\n**EXECUTE:**

\n
$$
\mathbf{A} = \lim_{t \to \infty} \mathbb{E} [\mathbf{A}_{t}] = \lim_{t \to \infty} \mathbb{E}_{\mu} \Big[F_{t} | S_{t} = s \Big]
$$
\nUsing $[\mathbf{f}]_{t} = d_{\mu}(1) = 0.5$

we have:

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

with
$$
[\mathbf{f}]_s = d_{\mu}(s) \lim_{t \to \infty} \mathbb{E}_{\mu}[F_t | S_t = s]
$$
 $[\mathbf{f}]_1 = d_{\mu}(1) = 0.5$
\nwe have:
\n $\mathbf{f} = \mathbf{d}_{\mu} + \gamma \mathbf{P}_{\pi}^{\top} \mathbf{d}_{\mu} + (\gamma \mathbf{P}_{\pi}^{\top})^2 \mathbf{d}_{\mu} + \cdots$
\n $= 0.5 + 0.9 \cdot 10$
\n $= 9.5$
\n $\mathbf{P}_{\pi} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

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

$$
\mathbf{F} \qquad \mathbf{I} - \gamma \mathbf{P}_{\pi} \qquad \text{key matrix}
$$

Figure 3: Emphatic TD approaches the correct value of zero, whereas conventional offpolicy TD diverges, on fifty trajectories on the $w \rightarrow 2w$ 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 $\lambda = 0$, $w_0 = 0$, $\alpha = 0.001$, 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