Aprendizaje Reforzado
Clase 5

Julián Martínez

Esquema "general"

$$v^{l+1}\left(S_t^{l+1}\right) = v^{l}\left(S_t^{l+1}\right) + \alpha\left[A\left(\underset{v^{l}}{\text{episodio } l+1},\right) - v^{l}\left(S_t^{l+1}\right)\right]$$

MC:

- No usa la markovianeidad
- precisa todo el episodio para actualizar la función de valor.
- Varianza.

TD:

- Usa el supuesto de Markovianeidad.
- Actualiza la función en cada paso del sampling.
- Para cada actualización utiliza la estimación previa de la función de valor (Sesgo).

GLIE control Monte Carlo

$$\lim_{k \to \infty} N_k(s,a) = \infty$$

$$\lim_{k \to \infty} \pi_k(a|s) = \delta_{\operatorname{argmax}_{a'} q^*(s,a')}$$

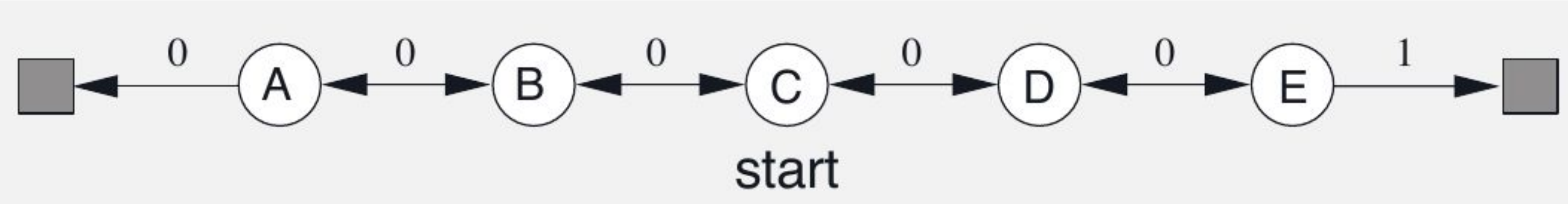
$$\varepsilon^{\ell} = \frac{1}{\ell}$$

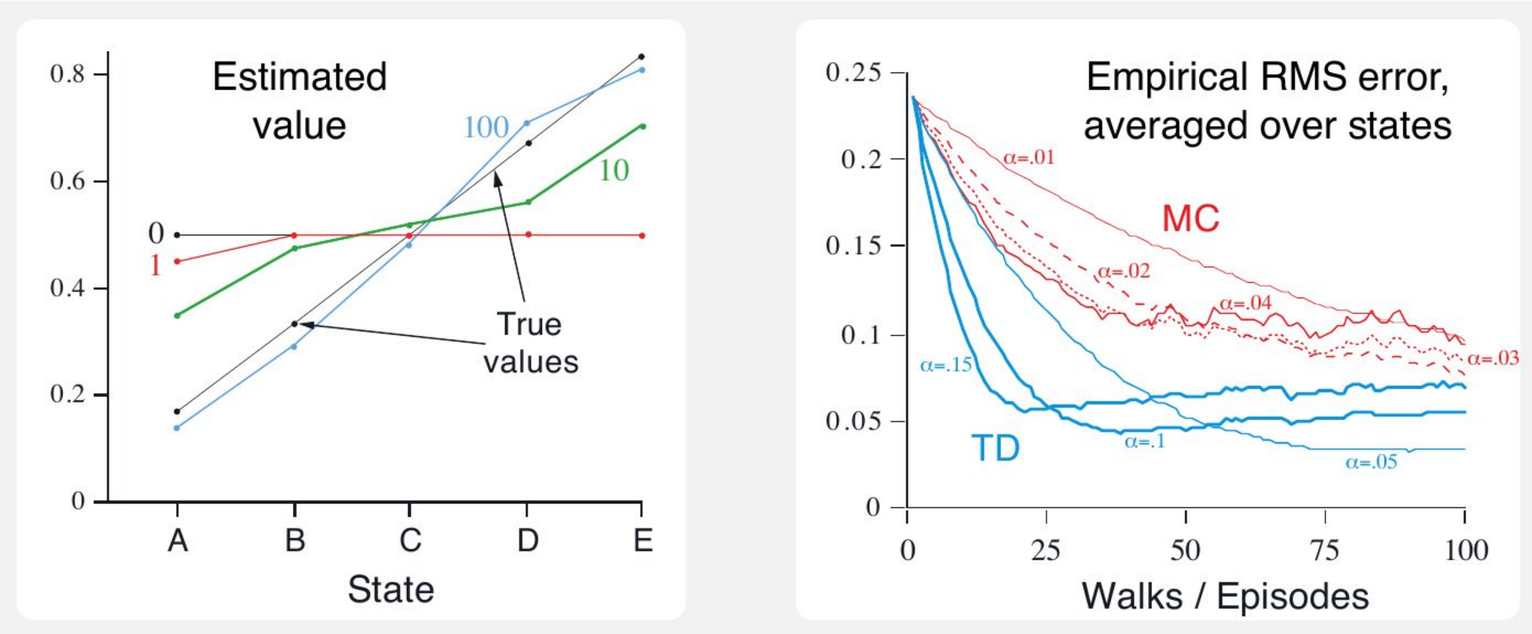
Pasa en las mejores familias

- The policy converges on a greedy policy,

$$\lim_{k \to \infty} \pi_k(a|s) = \mathbf{1}(a = \operatorname*{argmax}_{a' \in \mathcal{A}} Q_k(s, a'))$$

Random Walk Example 6.2 (Markov Reward Process)





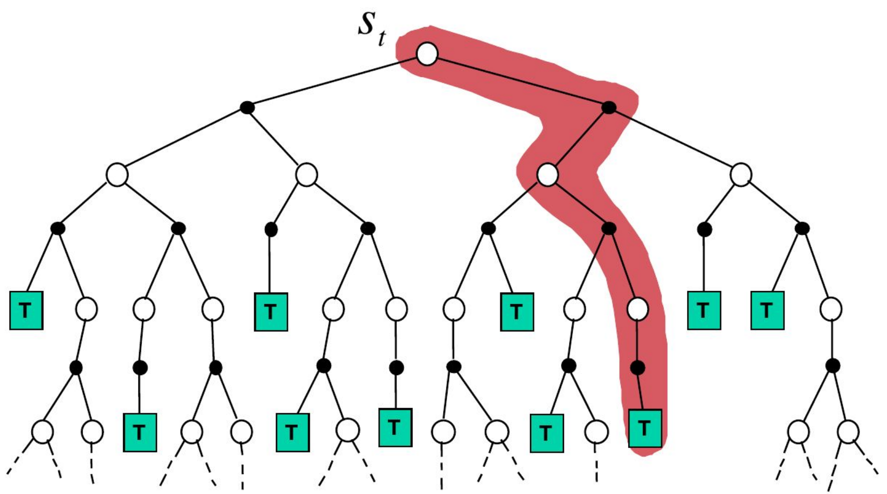
$$\sqrt{\sum_{r=1}^{100} \left\{ \frac{\sum_{s} \left[\hat{v}_r^{\ell}(s) - v(s) \right]^2}{|S|} \right\} \cdot \frac{1}{100}}$$

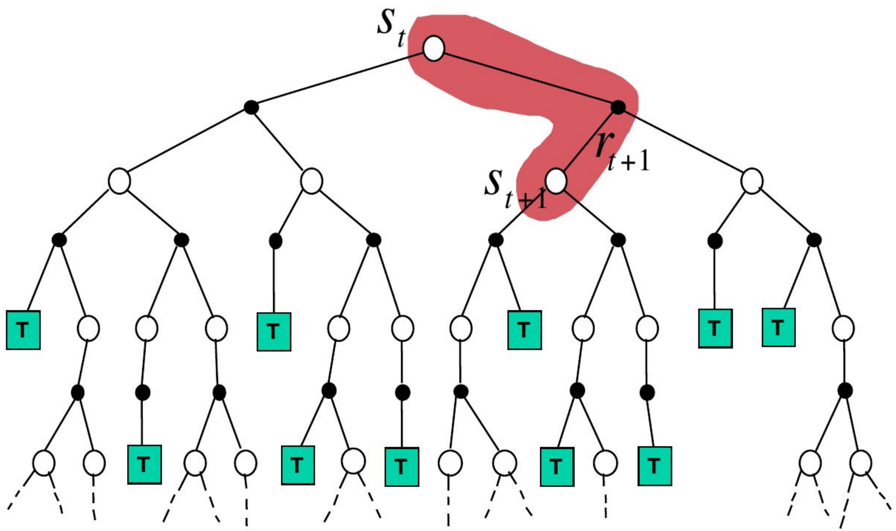
Relación entre los errores

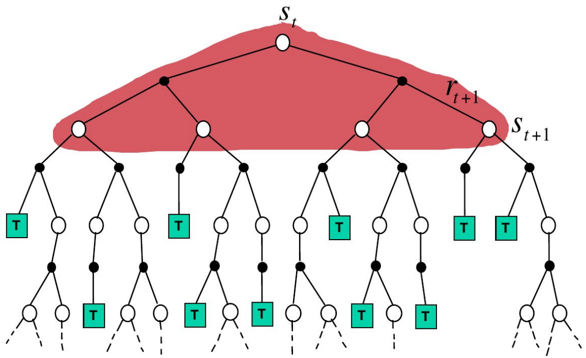
$$G_t^{k+1} - v^k(S_t) = \dots = \sum_{j=t}^{T-1} \gamma^{j-t} \delta_j^{TD}$$

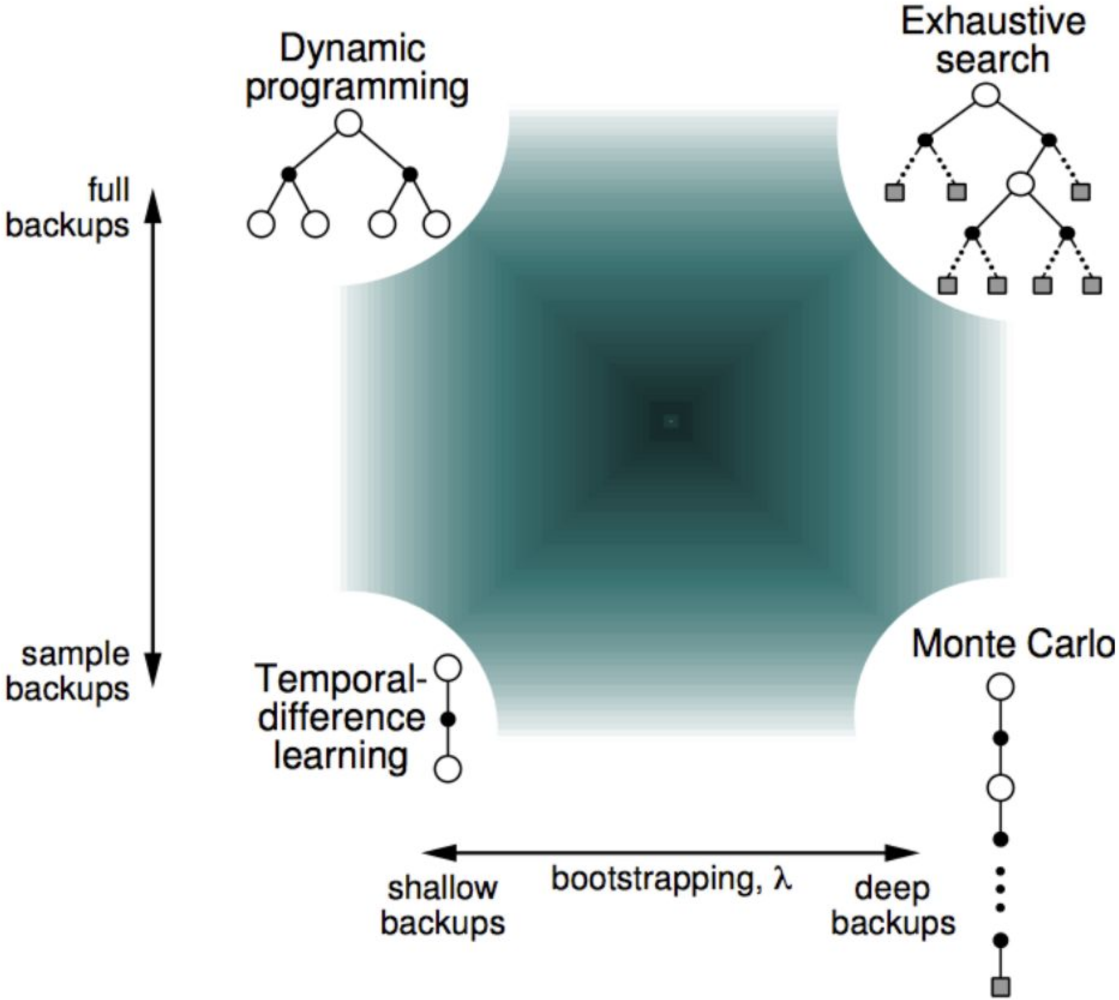
$$
\begin{aligned}
G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) && \text{(from (3.9))} \\
&= \delta_t + \gamma\big(G_{t+1} - V(S_{t+1})\big) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\big(G_{t+2} - V(S_{t+2})\big) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}\big(G_T - V(S_T)\big) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}\big(0 - 0\big) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k. && (6.6)
\end{aligned}
$$

Esquema de Backcups

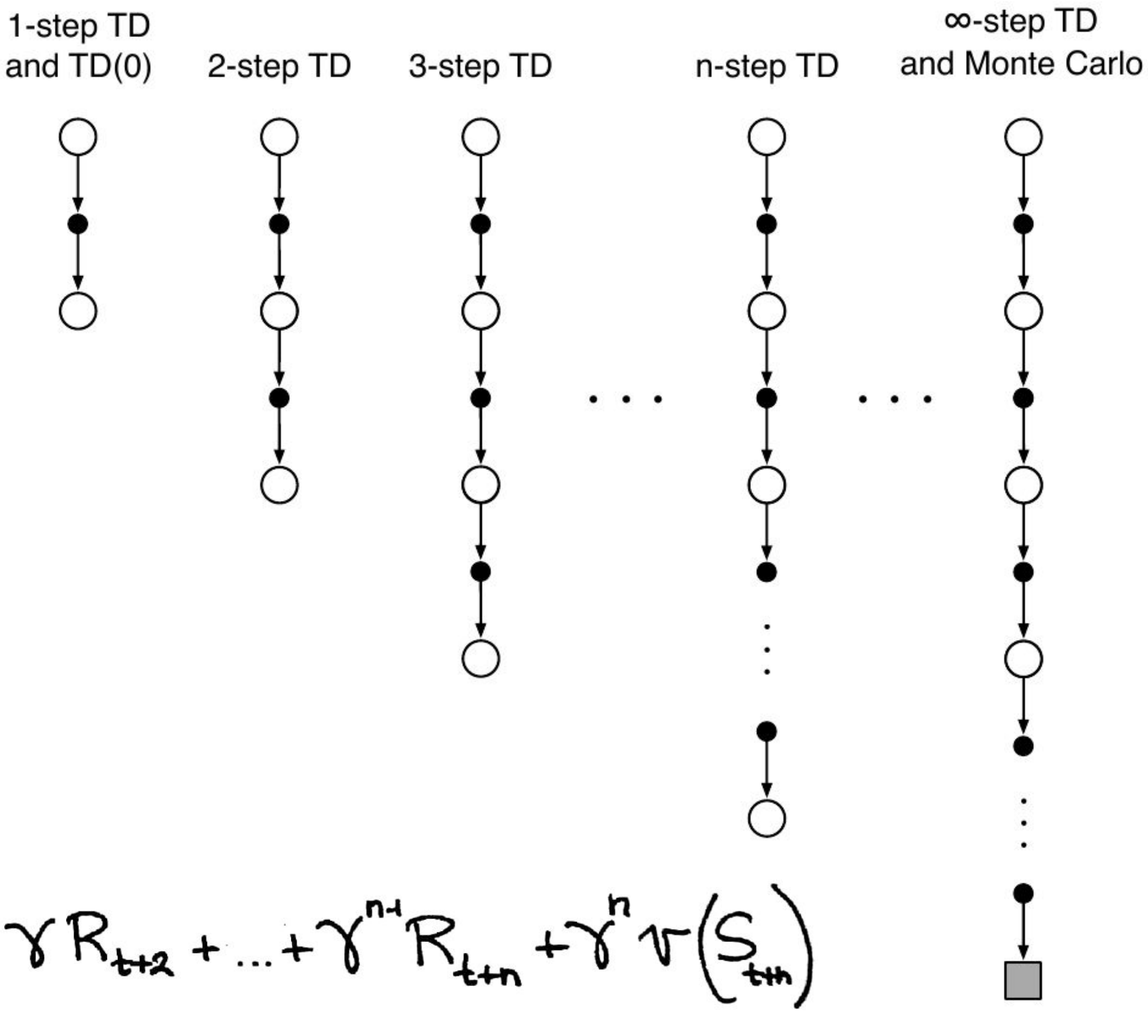







Dynamic programming
Exhaustive search
full backups
sample backups
Temporal-difference learning
Monte Carlo
shallow backups
bootstrapping, λ
deep backups

Entre MC y TD



$$G_{t:t+n} := R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^{n} v(S_{t+n})$$

N-STEP TD

$$G_t^{(n)} \quad G_{t:t+n} := R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n v(S_{t+n})$$

$$v^{t+n}(S_t) = v^{t+n-1}(S_t) + \alpha\left[G_{t:t+n} - v^{t+n-1}(S_t)\right]$$

Error reduction property

$$\max_s \left| \mathbb{E}_\pi\left[G_{t:t+n} \mid S_t = s \right] - v_\pi(s) \right|$$

$$\leq \gamma^n \max_s \left| V_{t+n-1}(s) - v_\pi(s) \right|$$

TD(λ) - Forma "ingeniosa" de mezclar los n-step

$$G_t^{\lambda} = (1-\lambda) \cdot \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

$$G_t^{(n)} = \sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^{n} \hat{v}(S_{t+n})$$

Observación "algorítmica"

$$\sum_{n=1}^{\infty}\left\{\sum_{k=1}^{n}\lambda^{n-1}\cdot\gamma^{k-1}\cdot R_{t+k}+\lambda^{n-1}\hat{v}(S_{t+n})\right\}(1-\lambda)$$

$$=\sum_{k=1}^{\infty}\sum_{n=1}^{k}(1-\lambda)\lambda^{n-1}\gamma^{k-1}R_{t+k}+(1-\lambda)\lambda^{n-1}\hat{v}(S_{t+n})$$

Esos pesos son la geométrica!

$$G_t^\lambda = (1-\lambda) \cdot \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

$$G_t^\lambda := \mathbb{E}_\lambda\left[G_t^{(M)}\right] \qquad M \sim \mathrm{Geo}(1-\lambda)$$

Que elegancia la de francia...

$$G_t = (1-\gamma)^{-1} \, \mathbb{E}_\gamma \left[R_{\tau+t} \right]$$

$$\tau \sim \mathrm{Geo}(1-\gamma)$$

Reinterpretando la función de valor

$$v(s) := \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

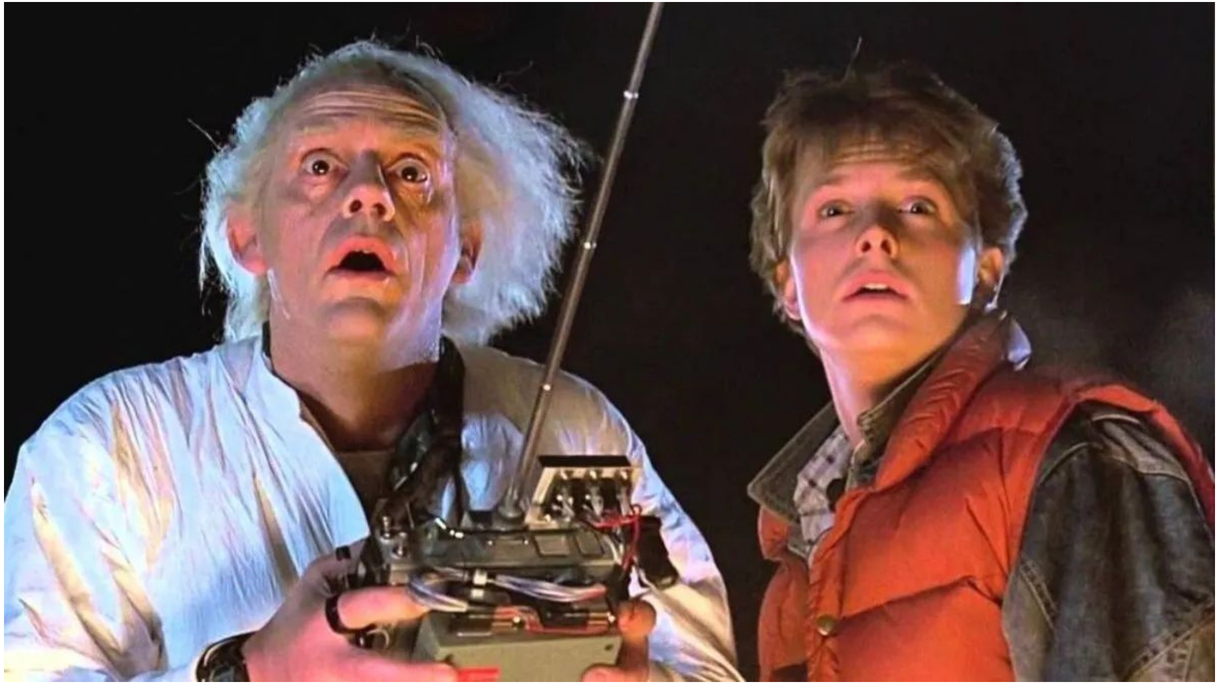
$$= (1-\gamma)^{-1} \, \mathbb{E}_\pi \left[\mathbb{E}_\gamma \left[R_{\tau+t} \right] \mid S_t = s \right]$$

$$E_\gamma\left[R_{\tau+t}\right] = E_\gamma\left[R_{\tau+t}\mathbb{1}_{\{\tau\leq n\}}\right] + P(\tau>n)\,E_\gamma\left[R_{\tau+t}\,\middle|\,\tau>n\right]$$

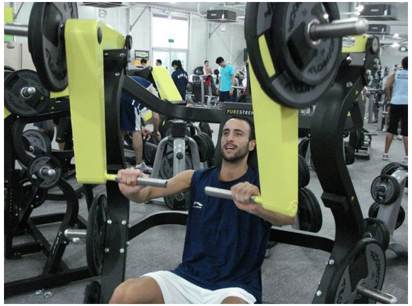
$$G_t^{(n)} = \mathbb{E}_\pi\left[E_\gamma\left[R_{\tau+t}\right]\,\middle|\,\mathcal{F}_n\right]$$

$$= E_\gamma\left[R_{\tau+t}\mathbb{1}_{\{\tau\leq n\}}\right] + \gamma^n v\left(S_{t+n}\right)$$

Pero TD(lambda) requiere de todo el episodio hasta el final...Tengo que mirar al futuro...



Eligibility Traces




¿qué explica mayormente el campeonato?

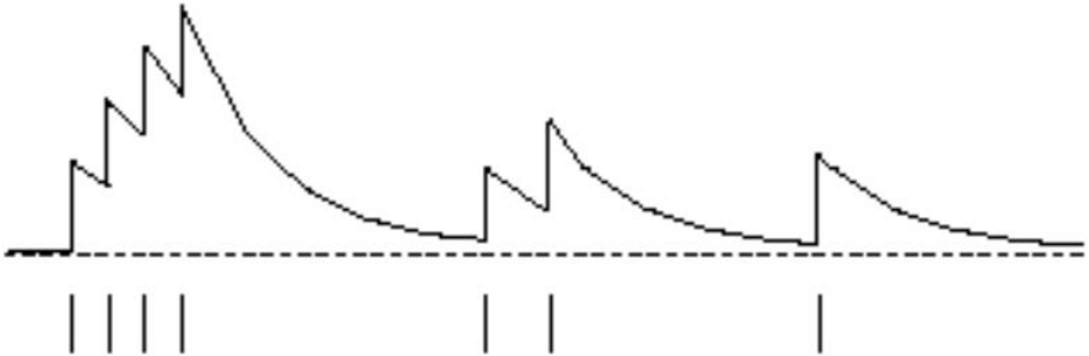
- El entrenamiento (frecuencia)
- Las vacaciones (cuán recientemente visité ese estado)

¿cómo combinar ambas?

Eligibility traces

$$E_0(s) = 0$$

$$E_t(s) = \gamma \lambda E_{t-1}(s) + \mathbf{1}(S_t = s)$$



$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

$$V(s) \leftarrow V(s) + \alpha \delta_t E_t(s)$$

Foward-view ~ Backward-view

Theorem

The sum of offline updates is identical for forward-view and backward-view TD(λ)

$$\sum_{t=1}^{T} \alpha \delta_t E_t(s) = \sum_{t=1}^{T} \alpha \left(G_t^\lambda - V(S_t) \right) \mathbf{1}(S_t = s)$$

El lado oscuro...

Off-policy Learning with Eligibility Traces: A Survey
https://hal.inria.fr/hal-00644516/PDF/jmlr.pdf

Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms

https://link.springer.com/content/pdf/10.1023/A:1007678930559.pdf


TENGO HASTA AHÍ...
GENERADORMEMES.COM