A Local Temporal Difference Code for Distributional Reinforcement Learning

Brabeeba Wang

4/30/2021

Traditional

$$V(s_t) \leftarrow V(s_t) + \alpha\delta(t)$$
$$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t)$$

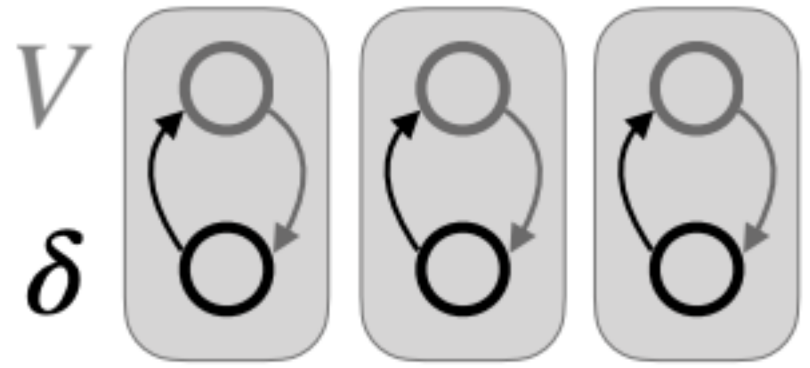
$$V(s) \rightarrow E\Big[\sum_{\tau=0}^{\infty} \gamma^{\tau} r_{t+\tau} \Big| s_t = s\Big]$$

Distributional

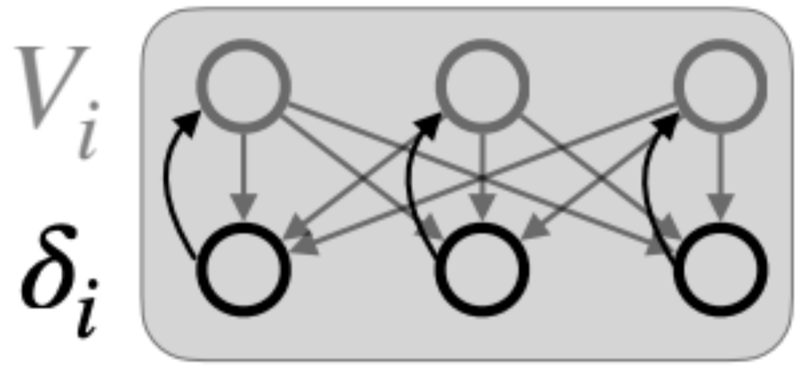
$$V_i(s_t) \leftarrow V_i(s_t) + \alpha_i^{+}\delta_i(t) \;\text{ if } \delta_i(t) > 0$$
$$V_i(s_t) \leftarrow V_i(s_t) + \alpha_i^{-}\delta_i(t) \;\text{ if } \delta_i(t) < 0$$
$$\delta_i(t) = r_t + \gamma\tilde{V}(s_{t+1}) - V_i(s_t)$$

But sampling from value distribution is not local.

(a)



(b)



Laplace code on discount factor

$$V_\gamma(s_t) \leftarrow V_\gamma(s_t) + \alpha\delta_\gamma(t)$$

$$\delta_\gamma(t) = r_t + \gamma V_\gamma(s_{t+1}) - V_\gamma(s_t)$$

Discrete:

$$V_\gamma(s_t) \to E\Big[\sum\nolimits_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau} \Big| s_t\Big] = \sum\nolimits_{\tau=0}^{\infty} \gamma^\tau E[r_{t+\tau}|s_t]$$

$$Z^{-1}\{V_\gamma(s_t)\}_{\gamma\in(0,1)} = \{E[r_{t+\tau}|s_t]\}_{\tau=0}^{\infty}$$

Continuous:

$$V_\gamma(s_t) \to \int_0^\infty e^{-\tau(-\log\gamma)}\, E[r_{t+\tau}|s_t]\, d\tau$$

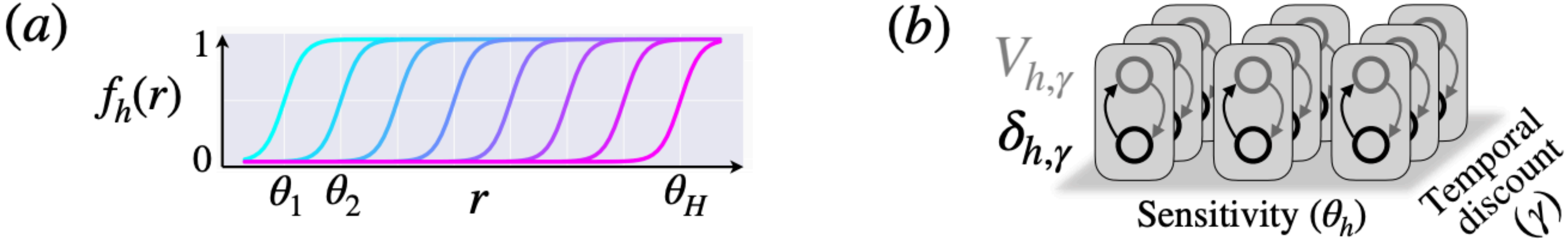
$$\mathcal{L}^{-1}\{V_\gamma(s_t)\}_{\gamma\in(0,1)} = \{E[r_{t+\tau}|s_t]\}_{\tau>0}$$

Linear readout: $\mathbf{L}^{-1}\big[V_{\gamma_1}(s_t), \ldots, V_{\gamma_\mathrm{N}}(s_t)\big] = \big[E[r_{t+0}|s_t], \ldots, E[r_{t+\mathrm{T}}|s_t]\big]$

Laplace code on reward sensitivity

$$V_{h,\gamma}(s_t) \leftarrow V_{h,\gamma}(s_t) + \alpha\delta_{h,\gamma}(t)$$

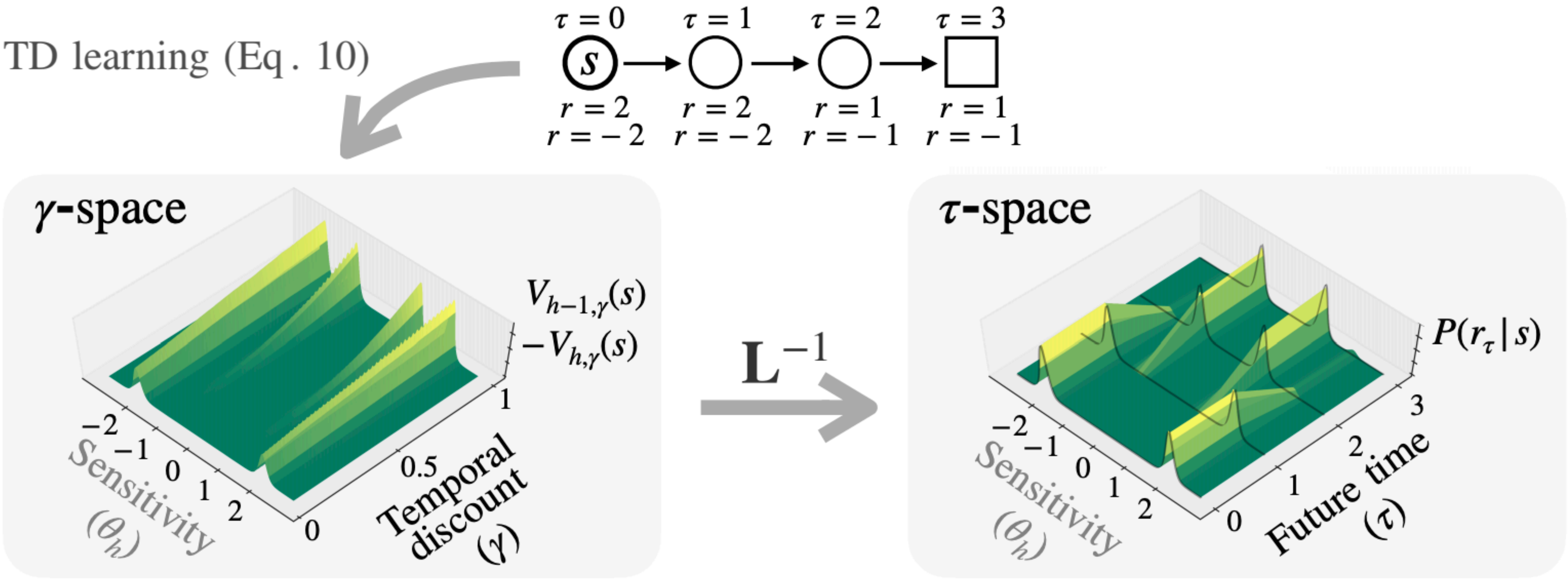
$$\delta_{h,\gamma}(t) = f_h(r_t) + \gamma V_{h,\gamma}(s_{t+1}) - V_{h,\gamma}(s_t)$$



$$V_{h,\gamma}(s_t) \rightarrow E\Big[\sum_{\tau=0}^{\infty} \gamma^\tau f_h(r_{t+\tau})\Big| s_t\Big] = \sum_{\tau=0}^{\infty} \gamma^\tau E[f_h(r_{t+\tau})|s_t]$$

$$= \sum_{\tau=0}^{\infty} \gamma^\tau E[H(r_{t+\tau} - \theta_h)|s_t] = \sum_{\tau=0}^{\infty} \gamma^\tau P(r_{t+\tau} > \theta_h|s_t)$$

$$\mathbf{L}^{-1}[V_{h,\gamma_1}(s_t), \ldots, V_{h,\gamma_\mathrm{N}}(s_t)] = [P(r_{t+0} > \theta_h|s_t), \ldots, P(r_{t+\mathrm{T}} > \theta_h|s_t)]$$



TD learning (Eq. 10)
τ = 0
τ = 1
τ = 2
τ = 3
s
r = 2
r = − 2
r = 2
r = − 2
r = 1
r = − 1
r = 1
r = − 1
γ-space
$V_{h-1,\gamma}(s)$
$-V_{h,\gamma}(s)$
Sensitivity (θ_h)
Temporal discount (γ)
$\mathbf{L}^{-1}$
τ-space
$P(r_\tau | s)$
Sensitivity (θ_h)
Future time (τ)

What can this code recover?

$$E\left[\sum_{\tau=0}^{\infty} \tilde{\gamma}^{\tau} r_{t+\tau} \middle| s_t\right] = \sum_{\tau=0}^{\infty} \tilde{\gamma}^{\tau} \sum_{r} r P(r_{t+\tau} = r | s_t)$$

$$E\left[\sum_{\tau=0}^{\infty} \tilde{\gamma}^{\tau} r_{t+\tau} \middle| s_t\right] = \sum_{h=1}^{H} \left(V_{h-1,\tilde{\gamma}}(s_t) - V_{h,\tilde{\gamma}}(s_t)\right)\theta_h$$

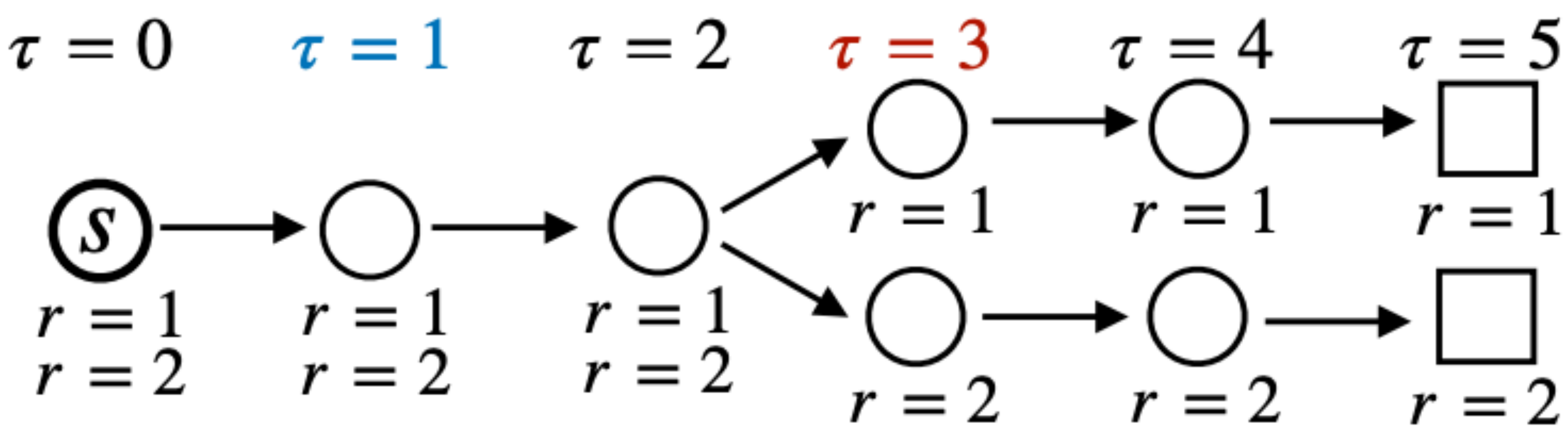
$$P\left(\sum_{\tau=0}^{T} \tilde{\gamma}^{\tau} r_{t+\tau} = V \middle| s_t\right) = (1 - \delta_{(V,0)}) \sum_{\tau=0}^{T} P\left(r_{t+\tau} = \tilde{\gamma}^{-\tau} V \middle| s_t\right)$$

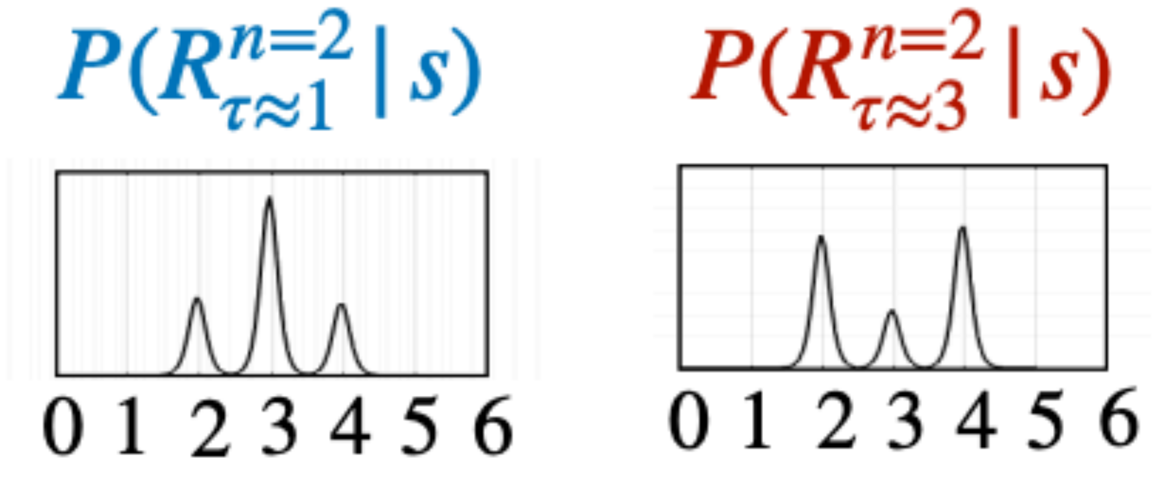
However, notice that this approach recovers the reward distribution evolution but not the value distribution since reward might be correlated across time.

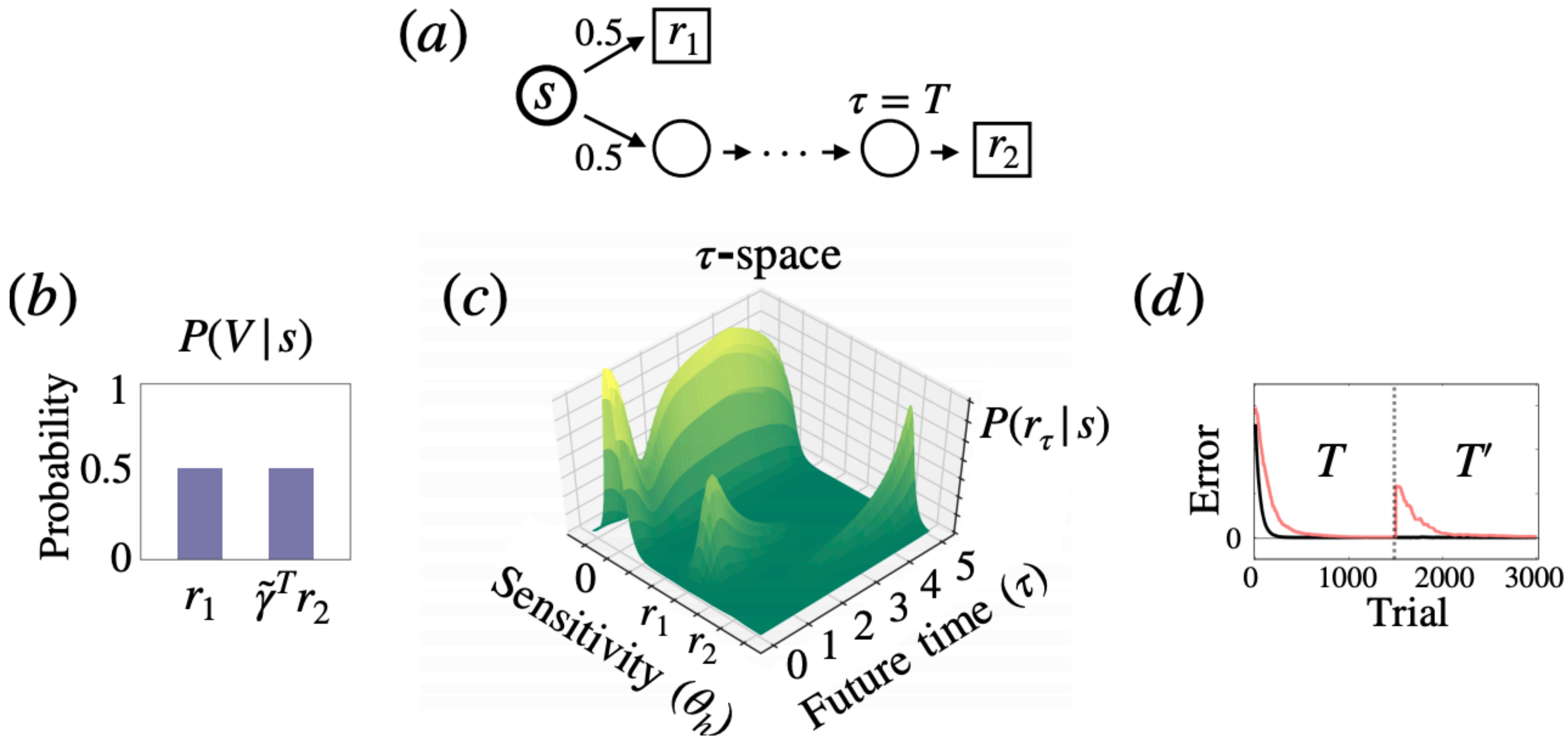
Laplace code on temporal discount

$$R_t^n = r_t + \ldots + \tilde{\gamma}^n r_{t+n}$$

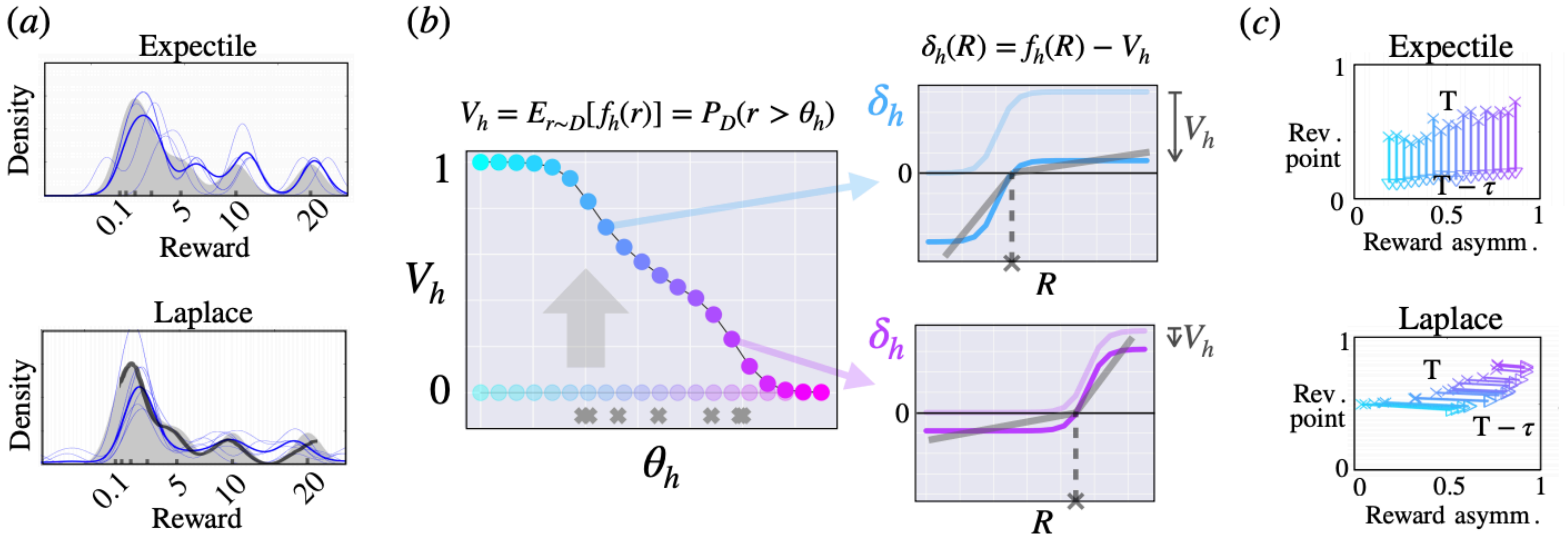
$$V_{h,\gamma,n}(s_t) \leftarrow V_{h,\gamma,n}(s_t) + \alpha \delta_{h,\gamma,n}(t+n)$$
$$\delta_{h,\gamma,n}(t+n) = f_h(R_t^n) + \gamma V_{h,\gamma,n}(s_{t+1}) - V_{h,\gamma,n}(s_t)$$







(a)
0.5
r_1
s
$\tau = T$
0.5
r_2
(b)
$P(V|s)$
Probability
1
0.5
0
r_1
$\tilde{\gamma}^T r_2$
(c)
τ-space
$P(r_\tau|s)$
0
r_1
r_2
Sensitivity (θ_h)
0 1 2 3 4 5
Future time (τ)
(d)
Error
0
T
T'
0
1000
2000
3000
Trial



(a)
Expectile
Density
0.1
5
10
20
Reward
Laplace
Density
0.1
5
10
20
Reward
(b)
$V_h = E_{r\sim D}[f_h(r)] = P_D(r > \theta_h)$
1
V_h
0
θ_h
$\delta_h(R) = f_h(R) - V_h$
δ_h
0
V_h
R
δ_h
0
V_h
R
(c)
Expectile
1
Rev. point
0
T
T − τ
0
0.5
1
Reward asymm.
Laplace
1
Rev. point
0
T
T − τ
0
0.5
1
Reward asymm.

Connection to successor representation

$$[SR^{\gamma}(s)]_{s'} = E\Big[\sum_{\tau=0}^{\infty} \gamma^{\tau} \delta_{(s', s_{t+\tau})} \Big| s_t = s\Big] = \sum_{\tau=0}^{\infty} \gamma^{\tau} P\big(s_{t+\tau} = s' | s_t = s\big).$$

$$V_{h,\gamma}(s_t) \to \sum_{\tau=0}^{\infty} \gamma^{\tau} \sum_{s} P\big(s_{t+\tau} = s | s_t\big) P\big(r_{t+\tau} > \theta_h | s_t, s_{t+\tau} = s\big)$$

$$V_{h,\gamma}(s_t) \to \sum_{s} \Big(\sum_{\tau=0}^{\infty} \gamma^{\tau} P\big(s_{t+\tau} = s | s_t\big)\Big) P\big(r > \theta_h | s\big) = SR^{\gamma}(s_t) \cdot \boldsymbol{r}_h$$