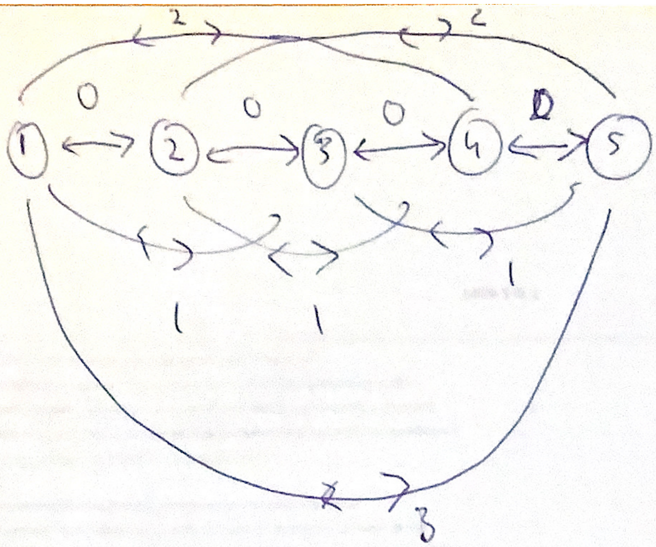
1 =



Exam
dec 2021
DPRL

a. $V_t(x) = \min_y \{ d(x, y) + V_{t-1}(y) \}$

$V_t(d) = 0 \ \forall t$ (d = destination

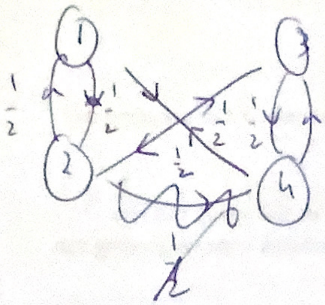
$V_0(x) = \infty \ \forall x \neq d$.

b. $V_t(x)$ is total minimal distance to d from x in max t steps

c

1	0	1	2	3	∞
2	0	0	1	2	∞
3	0	0	0	1	∞
4	0	0	0	0	∞
5	0	0	0	0	0
	4	3	2	1	0

2)



f no convergence if periodic transformation $P \to \delta P + (1-\delta) I$

a $\pi_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ $\pi_1 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}$, $\pi_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}$, $\pi_3 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}$, $\pi_4 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}$

b comm & periodic

because in 2 steps you can reach any state from any other &
periodic = 2, you alternate between 1/3 en 2/4

c $\begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$

d $\pi(1) = \frac{1}{2}\pi(2) + \frac{1}{2}\pi(4)$ $\qquad \sum \pi(x) = 1$

$\pi(2) = \frac{1}{2}\pi(1) + \frac{1}{2}\pi(3)$

$\pi(3) = \frac{1}{2}\pi(2) + \frac{1}{2}\pi(4)$ $\qquad \pi_* = \begin{pmatrix} \frac{1}{4} & & & \\ & \frac{1}{4} & & \\ & & \frac{1}{4} & \\ & & & \frac{1}{4} \end{pmatrix}$

$\pi(4) = \frac{1}{2}\pi(1) + \frac{1}{2}\pi(3)$

(one can be dropped)

e $V(1) + \phi = 1 + \frac{1}{2} V(2) + \frac{1}{2} V(4)$ $\qquad \phi = \frac{10}{4} = 2.5$

$V(2) + \phi = 2 + \frac{1}{2} V(1) + \frac{1}{2} V(3)$ $\qquad$ ~~set~~ $V(1) = 0$

$V(3) + \phi = 3 + \frac{1}{2} V(2) + \frac{1}{2} V(4)$ $\qquad V(3) = 2$

$V(4) + \phi = 4 + \frac{1}{2} V(1) + \frac{1}{2} V(3)$ $\qquad V(2) = 0.5$

$V(4) = 2.5$

3a) Recursion $Q_{t+1} = \gamma r_t + (1-\gamma) Q_t$ for some γ

r_t: 1 if Π, 0 if T, for each arm.

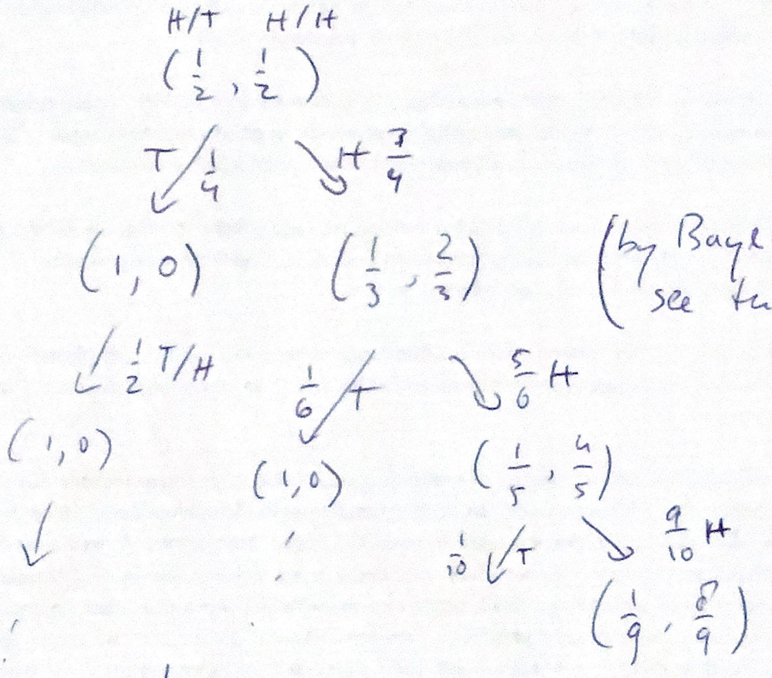
greedy: select highest Q_t

ε-greedy: highest with prob $1-\varepsilon$, random with ε

opt. values: start with $Q_0 = 1$, then greedy.

(or something higher, not lower)

3b)



(by Bayesian calc
see tutorial)

3c) GI: optimal exp. disc reward until stopping divided by exp. time until stopping

Stopping time: Time until stopping which depends on current state.

3d) pull any arm until T; if all arms have shown a T then select arbitrary arm.

Qa: Describe the difference between a model-based approach and a model-free approach (for the model-free approach, cite the two subfamilies of approaches).
A: Model-free is a direct approach and makes use of either value-based or policy-based learning. Model-based is an indirect approach that requires planning.

Qb. Similarly to supervised learning, the suboptimality (on the expected return) of an RL policy learned based on limited data can be decomposed into two terms. Cite these two terms and explain what they mean.
A: overfitting is the error term that drops to 0 if we have an infinite amount of data and the asymptotic bias is the bias introduced by the learning algorithm.

Qc. What are the two conditions for tabular Q-learning to converge (in the online setting)?
A: It converges w.p.1 to the optimal Q-function as long as
(i) $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$, and
(ii) the exploration policy π is such that $P_\pi [a_t = a l x_t = x] > 0, \forall (x, a)$.

Q_d. What is the specificity of the REINFORCE algorithm within the family of policy based methods
A: the Reinforce algorithm estimates the expected return with on-policy rollouts