

Value based

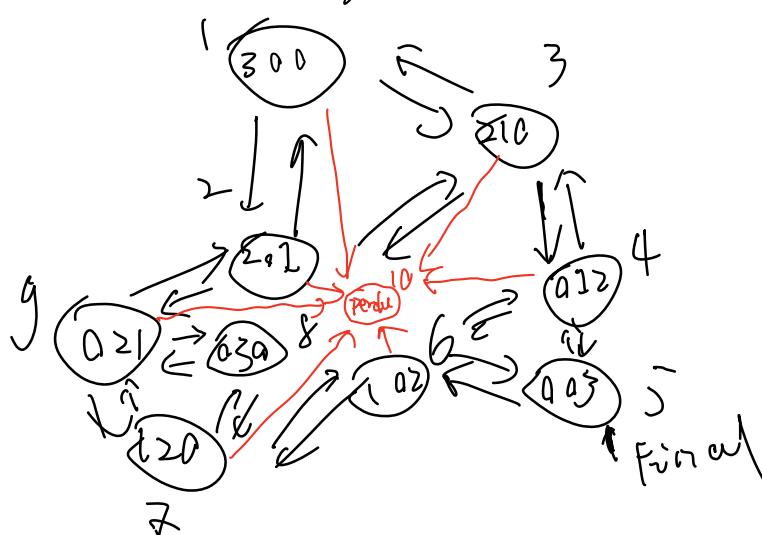
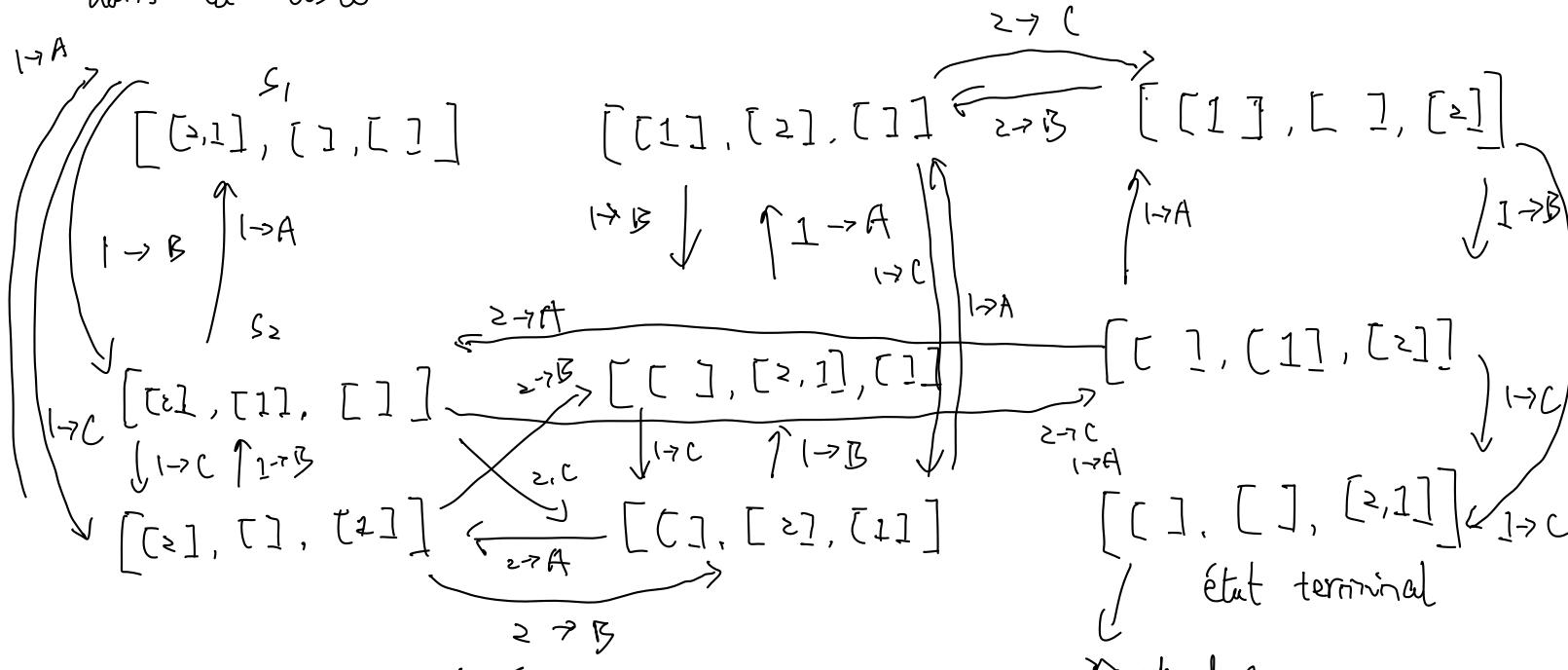
Tours de Hanoi

Q1.1. A, B, C. petit grand d1, d2 MDP

A B C
 $[[], [], []]$
 up down

$[[1, 2], [], []] \Rightarrow$

append quand le nombre est plus petit que la dernière élément dans la liste



Q1.2 Policy iteration, Value iteration.

policy iteration: $V_{i+1}^{\pi_k}(s) \leftarrow \sum_{s' \in S} P(s, \pi_k(s), s') [R(s, \pi(s), s') + \gamma V_i^{\pi_k}(s')]$

value iteration: $V_{i+1}(s) \leftarrow \max_{a \in A(s)} \sum_{s' \in S} P(s, a, s')$

- pas nombre infini de solution
- convergence

Q1.3 chaque opération $R(s, a, s') < 0$

arriver à l'état final $R(s, a, F) = 1$

Q1.4 $V^{\pi}_{t+1}(s) = \mathbb{E}_{\pi} (R_t | s_t = s) = \mathbb{E}_{\pi} \left[\sum_{i=0}^{+\infty} \gamma^i r_{t+i} \mid s_t = s \right]$

$V^{\pi}_{t+1}(s) = \sum_{a \in A(s)} \pi(s, a) (R(s, a, s') + \gamma V^{\pi}(s'))$
 $s' = \text{état arrivé de } (s, a)$

	1	2	3	4	5	6	7	8	9	R
V_0	1	1	1	1	1	1	1	1	1	{ -0,1 1 }
V_1	0,9	0,9	0,9	1,254	1,254	0,9	0,9	0,9	0,9	
V_2	0,8	0,8	0,8	0,8	0,974	0,974	0,974	0,974	0,8	

$\frac{2}{3} \times 0,9 + \frac{1}{3} \times (1+1)$

Q1.6
Sur un MDP, on a plus d'état final pour chaque état arrivé.



女口圖上

Q1.7 1, 3, 2, 9, 8, 7, 6, 4, 5

$$\gamma = 0.1$$

$R = -1000$ vers état perdu

1 Vers état 5
-1 Vers autre

éviter prendre le saut
plus présent

$R = -1$ Vers état 10

1 3 4 5

$R = 5$ Vers 5

$R = -1$ autre

$\gamma = 0, 1$

$R \geq 1000$ 10 = 1, 7, 2, 9, 7, 6, 4, 5

$R = 1$ 5 got discount, pas besoin de regard négatif
 $K = 0$ autre

$R = 0$
 $R = 1000$ very 5 1, 3, 4, 5
 $R = 0$ autre

Q 1.8

$TD(\lambda)$

si $\lambda = 0$ pas de variance
mais biais

$\lambda = 1$ pas de biais
bien plus de variance

Q 1.9 on corrige pas P et R

$TC(a) \rightarrow$ on peut pas faire at $A'(s)$, ou
argmax_{a $\in A'(s)$} $\sum_s P'(a, s)(R(a, s) + \gamma r(s))$

Q - learning

Q 1,10

R -1 $\Rightarrow 10$
 $\downarrow \alpha \alpha$ $\Rightarrow 5$ 1, 3, 4, 5 optimale
 $\alpha \Rightarrow$ audience
 $\downarrow \text{tais}$ étree une fois, on enlève
 1, 3, 2, 9, 8, 1, 6, 4, 5

Q 1,11

plus stable, plus lent

SARSA

Policy Gradient

Q. 2.1 

Q. 2.2.

consid^y, parti à partir de t.

Q. 2.3

réduire la variance

retire de puis rt une valeur qui dépend pas
d'action qu'on va prendre.

Q. 2.4

policy + value based

mais de variance : Q pour estimer le sr

inconvénient biaisé

↓
in reinforce c'est calculer
à la fin du trajet.

Q2.5

Sample plus important

Q2.6

Trajectory-based off-policy Deep Reinforcement Learning.

Q2.7

$$\nabla J(\theta) = \mathbb{E}_{\substack{\pi \sim d \\ T \sim \pi_L(T|s_0=a)}} \left[\sum_t^T r^+ Q^\pi(s_t, a_t) \nabla_\theta \log \pi(a_t | s_t) \right]$$

Q2.8:

DPG
SAC

GAW

?

WT F

Q. 3.1

GAW avec RL . Grille

a fort héritage réel pour discuter.
chercher les reward plus importants.

Q 3.2

Differences entre apprentissage et test
↳ simple manipulation de l'ensemble

GAW: explorer les espaces jamais explorés.

Q 3.3

Q 3.4

Q 3.5

Q 3.6

Q4.1

Q4.2

$$kL \quad 0$$

$$k=1 : PG$$

Q4.3

lige f

$$\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}$$

Q4.4

Q4.5

Q4.6: false done -

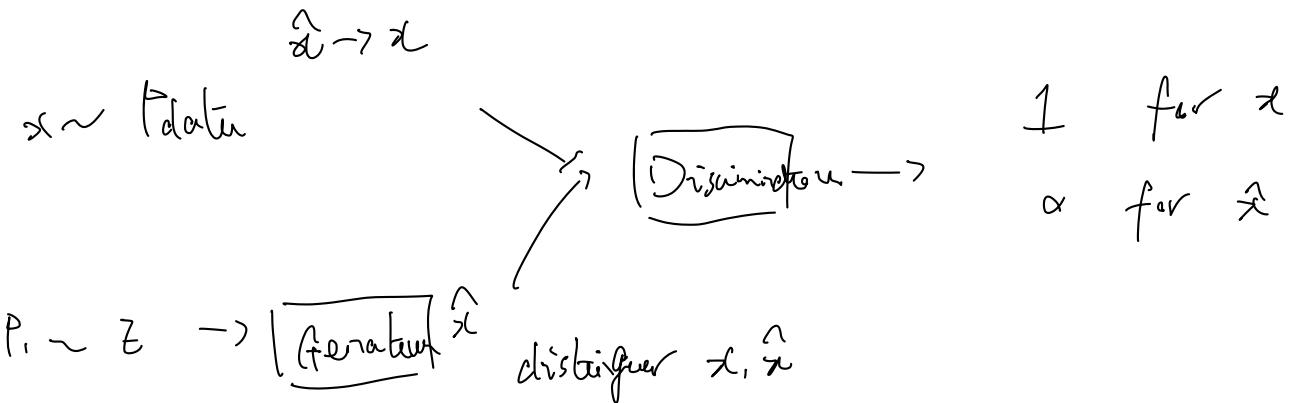
in ① agent iteration does observation
in ② return false done

Partie 2.

GAN

1.

- ① générateur : générer sample proche de vrai sample
- ② discriminateur : distinguer le vrai sample et sample généré



$$\min_{G} \max_{D} L(D, G) = \underbrace{\mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))]}_{\text{loss function}}$$

optimisation:

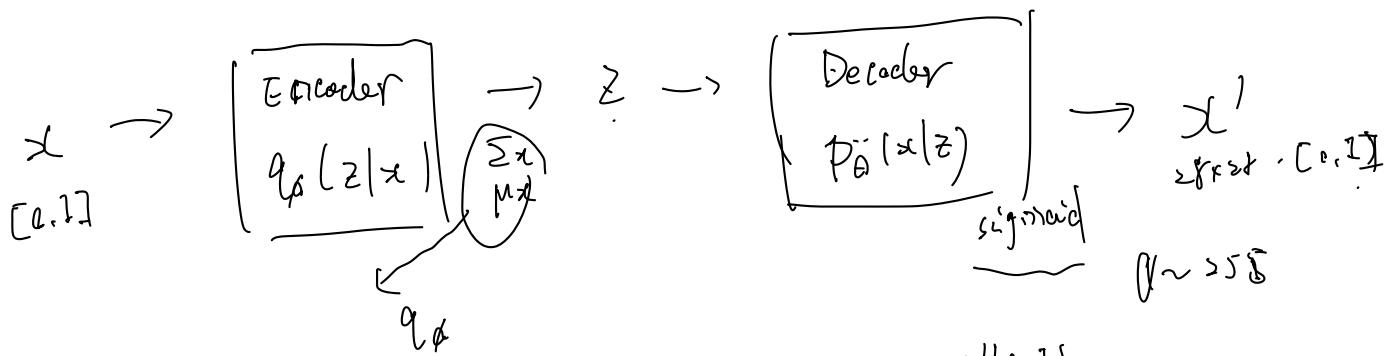
D:

$$\frac{1}{1 - D(G(z))}$$

$$G: \max_{z \sim p(z)} \mathbb{E} [\log D(G(z))] \quad (\text{steeper gradient})$$

Générateur clément

2, VAE

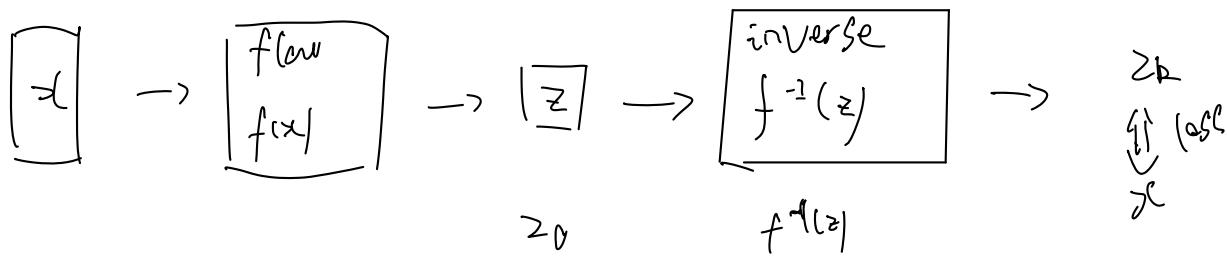


object: $\underbrace{V_L(\theta, \phi, x)}_{\text{lower bound}} = -\text{D}\text{KL}(q_\phi(z|x) \parallel p(z)) + \underbrace{E_{q_\phi(z|x)} [\log p_\theta(x|z)]}_{\text{regularization loss}} + \lambda \underbrace{\text{prior } N(0, I)}_{\text{cross entropy}}$

maximize object

Normalizing Flow

Distributions where sampling and density evaluation can be exact



Maximiser

$$\log p(x) = \log P_z(f_1^{-1}(x), \dots, f_k^{-1}(x)) - \sum_{i=0}^{k-1} (\log \left| \det \frac{\partial f_{i+1}}{\partial z_i} \right|)$$

$$\begin{aligned}
 P_x(x) &= P_z(g(x)) \cdot \left| \det \left(\frac{\partial g(x)}{\partial x} \right) \right| \quad g = f^{-1} \\
 &= P_z(F^{-1}(x)) \cdot \left| \det \left(\frac{\partial F}{\partial x} \right) \right|^{-1}, \quad F = f_0 \circ \dots \circ f_k \\
 &\quad F^{-1} = f_k^{-1} \circ \dots \circ f_0^{-1}
 \end{aligned}$$

$$\begin{aligned}
 &= P_z(f_k^{-1}(x), \dots, f_0^{-1}(x)) \cdot \prod_{i=0}^{k-1} \left| \det \frac{\partial f_{i+1}}{\partial z_i} \right|^{-1} \\
 &= \log P_z(f_k^{-1}(x), \dots, f_0^{-1}(x)) - \sum_{i=0}^{k-1} \log \left| \det \frac{\partial f_{i+1}}{\partial z_i} \right|^{-1}
 \end{aligned}$$

Rejet

implément

GAN: pas distribution dans donnée

VAE: existence de distribution dans la donnée

NF: pas de proximation, inversible.

1. oui. minimiser la borne inf

2. Oui

3. ?

4. Oui

5. Non

6.

7, oui

8, Non

9. Oui

10. Non

11. Non

12. Oui

13. Non, cross-entropy ou oui ?

14, oui

15, oui

16, oui ?

17, oui

18, Non ?

19, oui

20, oui

21, ?

22, oui ?

23, oui

24, Non veute

25,

↑

↓

○

↖