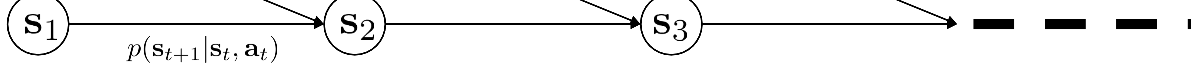
Dpto. Informática e Ingeniería de Sistemas.


$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_3$
$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

transition model

some applications (e.g.: inverse RL)

$$\pi(x) \leftarrow \arg\max_{a\in\mathcal{A}} \sum_{x'\in\mathcal{X}} p(x'|x,a)(R(x,a,x') + \gamma V_k(x'))$$

Universidad Zaragoza 1542

which actions are likely *a priori*
(assume uniform*)

*passive dynamics: non-uniform a priori actions can be incorporated in the "observation"

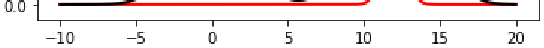
$$V_t(\mathbf{s}_t) \rightarrow \max_{\mathbf{a}_t} Q_t(\mathbf{s}_t, \mathbf{a}_t) \text{ as } Q_t(\mathbf{s}_t, \mathbf{a}_t) \text{ gets bigger!}$$

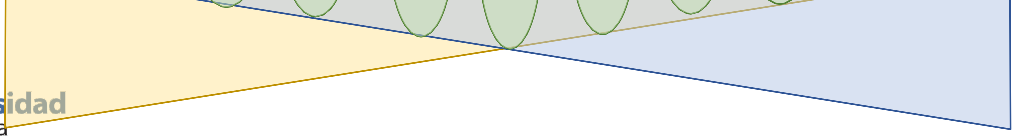
$$= \frac{p(\mathcal{O}_{t:T}|\mathbf{a}_t, \mathbf{s}_t)}{p(\mathcal{O}_{t:T}|\mathbf{s}_t)} \frac{p(\mathbf{a}_t, \mathbf{s}_t)}{p(\mathbf{s}_t)} = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)} \cancel{p(\mathbf{a}_t|\mathbf{s}_t)}$$

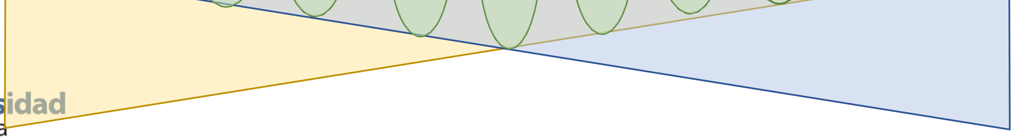
$$\boxed{\pi(\mathbf{a}_t|\mathbf{s}_t) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}}$$

Universidad Zaragoza 1542

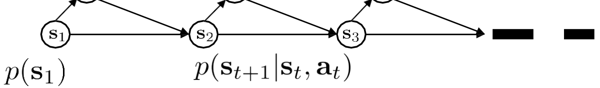
Universidad Zaragoza 1542


0.0
−10
−5
0
5
10
15
20





“given that you obtained high reward, what was your transition probability?”



Universidad Zaragoza 1542

$$KL(q||p) = \int_x q(x) \log \frac{q(x)}{p(x)} dx$$

- Minimizing KL-divergence maximizes expected log p and q entropy

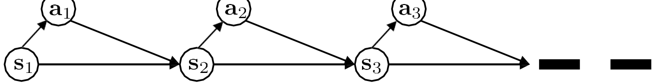
Universidad Zaragoza 1542

$$p(y) = \int_x p(y|x)p(x)dx = \int_x p(y, x)dx$$

- When we maximize the ELBO, we minimize the KL divergence!

$$\log E[y] \geq E[\log y]$$

$q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$



$p(\mathbf{s}_1)$ $\quad p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

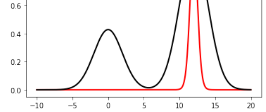
$$= \sum_t E_{(\mathbf{s}_t, \mathbf{a}_t) \sim q} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \mathcal{H}(q(\mathbf{a}_t | \mathbf{s}_t)) \right]$$

Universidad Zaragoza 1542

maximize reward and maximize action entropy!

$$V(\mathbf{s}_T) = \log \int \exp(Q(\mathbf{s}_T, \mathbf{a}_T)) d\mathbf{a}_T$$

$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t$$



Universidad
Zaragoza
1542

- Published multiple times between 2017-2018!

Haarnoja*, Tang*, 2017

3. Interact with the world, collect more data

Haarnoja, Zhou, Hartikainen, Tucker, Ha, Tan, Kumar, Zhu, Gupta, Abbeel, L. Soft Actor-Critic Algorithms and Applications. ‘18

[redacted]

1542

- Csaba Szepesvari, Algorithms of Reinforcement Learning
  - https://sites.ualberta.ca/~szepesva/rlbook.html