RL seminar #2: Imitation learning and policy gradient theorem

Maksim Kretov

MIPT, Deep learning lab & iPavlov.ai

28 Oct 2017

Outline

Class information

- Assignments

RL introduction

- Directed graphical models

Questions

- Imitation learning
- Policy gradient
- Quiz-related questions

Next steps

Table of Contents

Assignments

Coding

Deadline: 2 Nov 2017 (Thursday)
1 submission

Quiz

Deadline: 26 Oct 2017
24 submissions: rating will be prepared next week

Questions

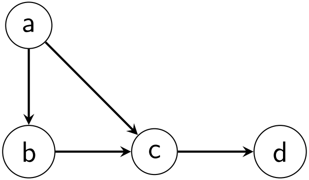
Few questions.

Table of Contents

RL introduction

Directed graphical models

aka Belief networks or Bayesian networks.



Factorization of complex joint distribution into simpler conditional probability distributions:

$$p(a, b, c, d) = p(a)p(b|a)p(c|a, b)p(d|c) = \prod_i p(x_i|Par_G(x_i))$$

Assumptions about which variables are conditionally independent from each other.

Exponential gain in number of parameters: $O(N^4)$ vs $O(N^3)$

RL introduction

D-separation (dependence separation)

Variables a and b are not separated, if they are connected by a path involving only unobserved variables.

Restrictions

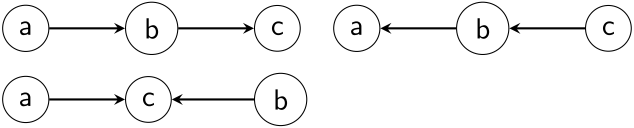
Context-specific independences are not possible to represent with graphical notations.

Example

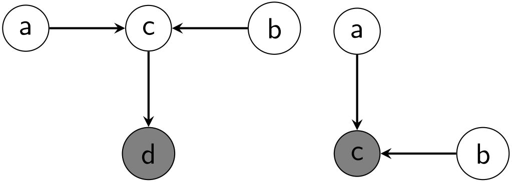
Three binary variables: a, b, c. When $a = 0$ then b and c are independent. But when $a = 1$, deterministically $b = c$. Using graphical notations, we cannot indicate that b and c are independent when $a = 0$.

RL introduction

Active paths between a and b (no d-separation)



Common cause for a and b



V-structure (explaining away) for a and b

$\Rightarrow$ When we observe node d or c in last 2 cases, we activate path between a and b and they are no longer d-separated.

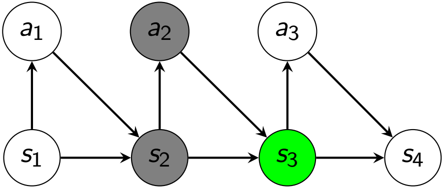
RL introduction

Markovian property

$$p(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}..) = p(s_{t+1}|s_t, a_t)$$

Future depends only on the present and doesn't depend on the past.

Using graphical models' terms:



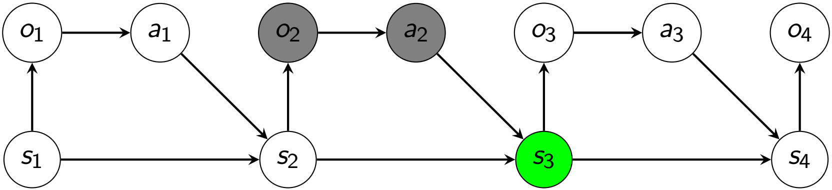
RL introduction

Partially observable environment (POMDP)

We have access only to observations o_t:

$$p(s_{t+1}|o_t, a_t, o_{t-1}, a_{t-1}..) \neq p(s_{t+1}|o_t, a_t)$$

Path from o_1 to s_3 is active:



$\Rightarrow$ Have to take into account full history of observations.

Table of Contents

Imitation learning: additional remarks

Multimodal behavior

Agent performs different actions given (almost) the same history. All actions are reasonable. Supervised learning can fail.

Implicit density models

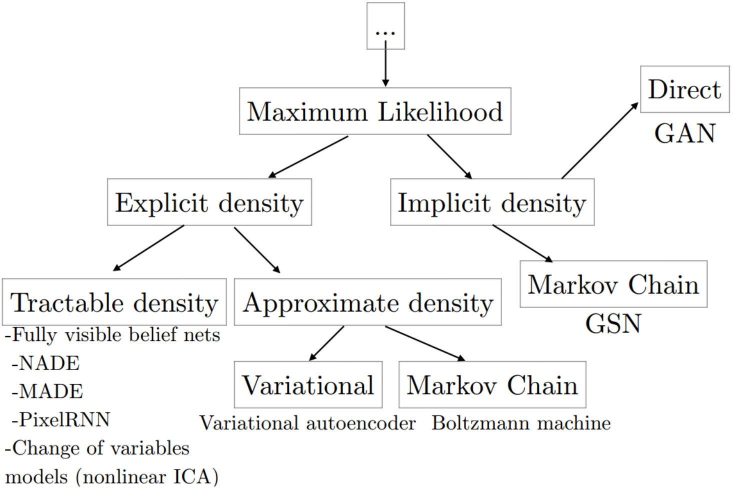
Examples: SGNs, GANs. They are called **implicit** because we do not model probability distribution directly.
Implicit distributions are (usually) intractable distributions from which we can easily sample and calculate gradient of expectations w.r.t. model parameters.

Alternative

Compare this with Gaussian policy, where we directly model mean and variance of the distribution and then sample.

Imitation learning: additional remarks

Taxonomy of deep generative models[1]



[1] Scheme from: https://arxiv.org/pdf/1701.00160.pdf

Imitation learning: additional remarks

Autoregressive discretization

Remedy for huge action space.

Procedure

If $dim(A) = N$, we introduce N models for sequential generation of actions. Then apply usual supervised learning procedure for components of action vector.

Why this can in principle work?

Because it just uses factorization of joint distribution according to Bayes' rule:

$$p(a_1, ..a_n) = \prod_{k=1}^{n} p(a_k | a_1, ..a_{k-1})$$

And idea is to approximate every factor above by a separate model.

Policy gradient: additional remarks

REINFORCE rule

Score function estimator for gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_\tau [\sum_t (R_t - b) \nabla_\theta \log \pi_\theta(a_t|s_t)]$$

Control variates (baselines)

$$\mathbb{E}\left[f(x)\right] \to \frac{1}{k} \sum_i (f(x_i) - \mu g(x_i)) - \mu \mathbb{E}\left[g(x)\right]$$

Extreme case: zero variance in case we already solved the task: $g = f, \mu = 1$.

Policy gradient: additional remarks

Intuitive example of high variance[2]

$$\log p_\theta(t|x,z) = \begin{cases} -100, & \text{with probability } 0.5 \\ -110, & \text{with probability } 0.5 \end{cases} \qquad \text{Mean} = -105,\ \text{Var} = 25$$

$$\nabla_\theta \log p_\theta(z|x) = \begin{cases} 1, & \text{with probability } 0.5 \\ -1, & \text{with probability } 0.5 \end{cases} \qquad \text{Mean} = 0,\ \text{Var} = 1$$

$$\log p_\theta(t|x,z) \nabla_\theta \log p_\theta(z|x) = \begin{cases} 110, & \text{with probability } 0.25 \\ 100, & \text{with probability } 0.25 \\ -100, & \text{with probability } 0.25 \\ -110, & \text{with probability } 0.25 \end{cases} \qquad \text{Mean} = 0,\ \text{Var} = 11050$$

$$(\log p_\theta(t|x,z) - c) \nabla_\theta \log p_\theta(z|x) = \begin{cases} 5, & \text{with probability } 0.5 \\ -5, & \text{with probability } 0.5 \end{cases} \qquad \text{Mean} = 0,\ \text{Var} = 25$$

$$c = -105$$

[2] Slide from Mikhail Figurnov's lecture on Deep Bayes 2017.

Discussion: questions from quiz

Tasks that do not fit RL framework

For discussion: action space is unknown or it is changing.

Non-standard application of RL

Meta-learning.

Any other questions?

Table of Contents

Next steps

Plan for the week

- Quiz N2: 1 Nov (Wednesday)
- Rating is coming: 3 Nov (Friday)
- Home assignment N2: 8 Nov (Wednesday)

Reading

Lectures 5-8 of CS294.
Please, post your questions about lectures in google doc:
`https://goo.gl/qN6jmJ`