Learning to behave via Imitation

ESSAI 2024 Course
Lecture 4/5

George Vouros

University of Piraeus, Greece

July 25, 2024

Outline

- Day 1: Motivation & Introduction to Deep Reinforcement Learning
- Day 2: Inverse Reinforcement Learning and Connections to Probabilistic Inference
- Day 3: Imitation Learning
- **Day 4: Non-Markovian, Multimodal Imitation Learning**
- Day 5: Imitating in Constrained Settings, Multiagent Imitation Learning.

Imitation Learning

Problem (ambiguous) statement

Given a set of demonstrated trajectories D generated by an unknown expert policy π_ϵ, learn a policy π that generates trajectories that are "as close as possible" to the expert trajectories.

Imitation learning

What can go wrong?

- Lack of training data
- Noisy or erroneous training data
- Distribution mismatch
- Compounding errors
- Discrimination ability (different actions in very similar settings)

Imitation Learning

What else can go wrong?

- Partial observability imposing non-Markovian behaviour
- Collapsing multi-modal behaviour in executing tasks in a single policy

Imitation Learning

non-Markovian Behaviour

$$\pi_\theta(a_t|o_t)$$

vs

$$\pi_\theta(a_t|o_1, o_2, ..., o_t)$$

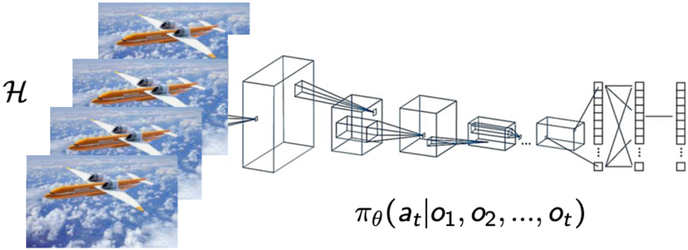
Usually behaviour depends on history of observations:

$$\pi_\theta : \mathbf{H} \to \mathcal{P}(\mathcal{A})$$

where $\mathbf{H} = \prod_{i=1}^{t} O,\ t = 2, 3....$
History provides **(temporal) context**.

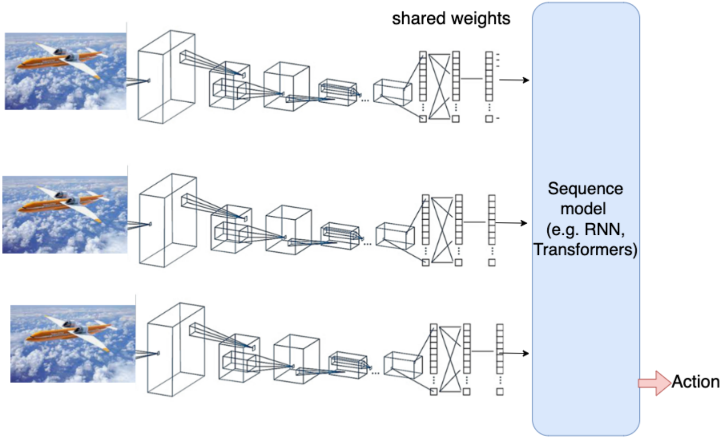
Imitation Learning

non-Markovian Behaviour: Basic



Imitation Learning

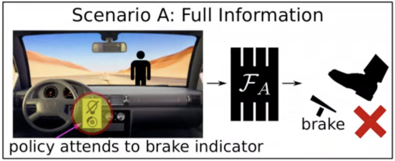
non-Markovian Behaviour: with Sequential models



Imitation Learning

non-Markovian Behaviour: with Sequential models

Using **H** may exacerbate correlations occurring in demonstrations: Instantiations of an action correlate to future actions.

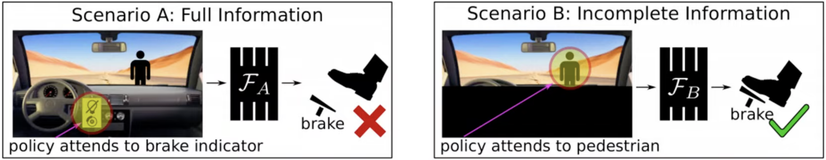


See also de Haan et al., "Causal Confusion in Imitation Learning"

Imitation Learning

non-Markovian Behaviour: with Sequential models

Causal misidentification: access to more information leads to worse generalization performance in the presence of distributional shift.



See also de Haan et al., "Causal Confusion in Imitation Learning"

Imitation Learning

non-Markovian Behaviour: with Sequential models

Using **H** may exacerbate correlations occurring in demonstrations: Instantiations of an action correlate to future actions.

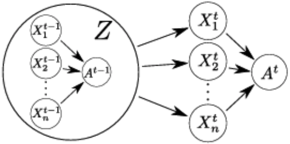


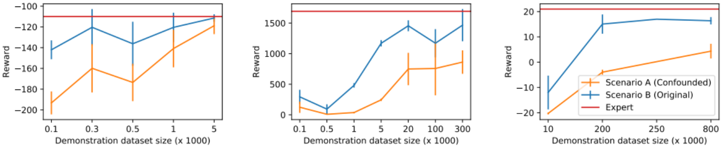
Figure 2: Causal dynamics of imitation. Parents of a node represent its causes.

See also de Haan et al., "Causal Confusion in Imitation Learning"

Imitation Learning

non-Markovian Behaviour: with Sequential models

Causal misidentification: access to more information leads to worse generalization performance in the presence of distributional shift.



(a) MountainCar (b) Hopper (c) Pong

Figure 4: Diagnosing causal misidentification: net reward (y-axis) vs number of training samples (x-axis) for ORIGINAL and CONFOUNDED, compared to expert reward (mean and stdev over 5 runs). Also see Appendix E.

See also de Haan et al., "Causal Confusion in Imitation Learning"

Imitation Learning

non-Markovian Behaviour: with Sequential models

Using **H** may exacerbate correlations occurring in demonstrations: **Causal misidentification** is the phenomenon whereby cloned policies fail by misidentifying the causes of expert actions

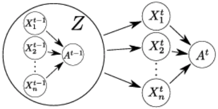


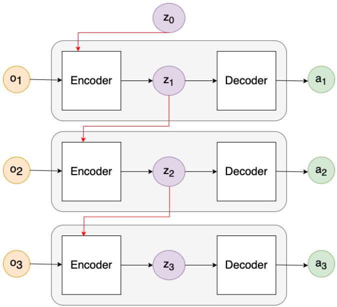
Figure 2: Causal dynamics of imitation. Parents of a node represent its causes.

Solutions proposed:

- Learn policies corresponding to *various* causal graphs
- Perform targeted interventions to efficiently search over the hypothesis set for the correct causal model.
    - Intervention with expert advise (DAgger style)
    - Use environmental returns (if you can) and compute the likelihood of graphs by means of $exp(R)$, rolling-out the policies

Imitation Learning

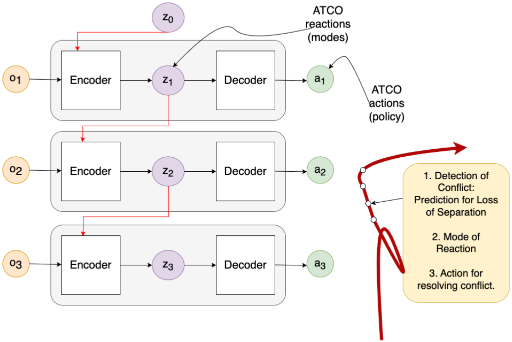
non-Markovian Behaviour: with VAE



See also Bastas and Vouros “Data-driven prediction of Air Traffic Controllers reactions to resolving conflicts”

Imitation Learning

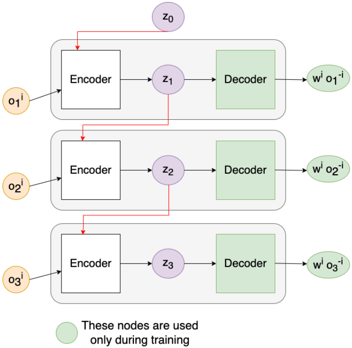
non-Markovian Behaviour: with VAE in a supervised setting



See also Bastas and Vouros "Data-driven prediction of Air Traffic Controllers reactions to resolving conflicts"

Imitation Learning

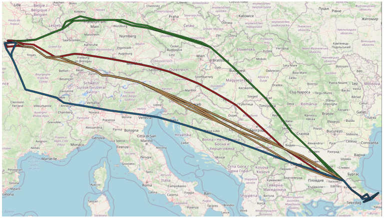
non-Markovian Behaviour: with VAE for state reconstruction (although originally proposed in a MARL setting)



Work done by A.Kontogiannis et al.

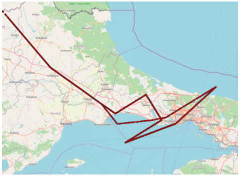
Imitation Learning

Multimodal Behaviour



Imitation Learning

Multimodal Behaviour



Imitation Learning

Multimodal Behaviour

The same expert may take different actions in the same situation.

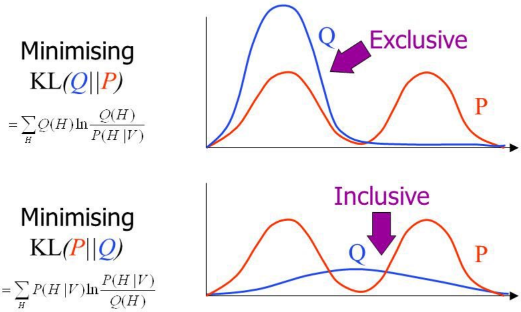
Imitation Learning

Multimodal Behaviour: Mode collapse

Most imitation learning algorithms suffer from mode collapse: I.e. their inability to distinguish between modalities and learn the average.

Multimodal Behaviour: Mode collapse

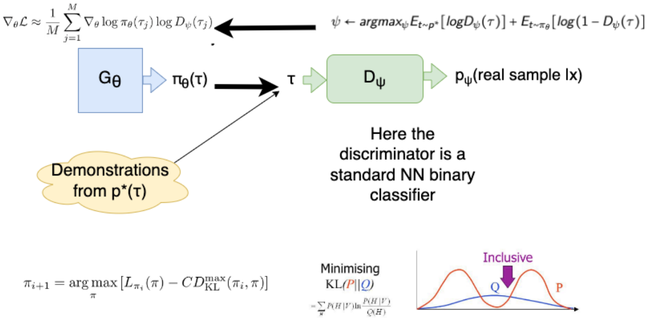
Given the analogy, take D_{KL} as an example: Given that P is the state action distribution from the demonstrations, and Q is the state action distribution learnt.
Should you compare Q against P or P against Q?



Imitation Learning

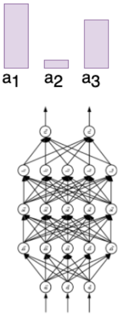
Multimodal Behaviour: Mode collapse

GAIL suffers from the mode collapse problem.



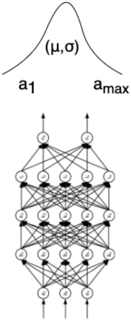
Imitation Learning

Multimodal Behaviour for discrete actions



Imitation Learning

Multimodal Behaviour for continuous actions



Averaging different modalities !

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
- Discretization with high-dimentional action spaces
- Compute the likelihood of each different option (and break ties randomly)

Multimodal Behaviour for continuous actions

How to resolve this problem?

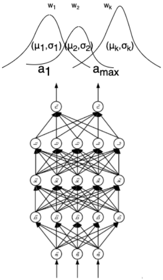
- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - Diffusion models

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - **Mixture of Gaussians**



$$\pi(a|o) = \sum_i w_i \mathcal{N}(\mu_i, \sigma_i)$$

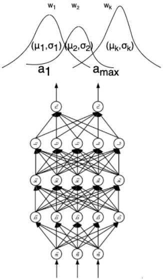
  - Latent variable models
  - Diffusion models

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - **Mixture of Gaussians**



$$\pi(a|o) = \sum_i w_i \mathcal{N}(\mu_i, \sigma_i)$$

You must choose k: number of modes (how many?)

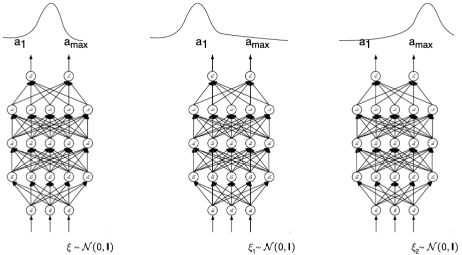
  - Latent variable models
  - Diffusion models

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models**
    Using latent variables models you can represent any distribution (conditional to the size of the NN).



    The particular way to correlate these variables to actual inputs/outputs is by means of variational autoencoders (VAEs).
  - Diffusion models

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The InfoGAIL case**

Li et al 2017, "InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The InfoGAIL case**

InfoGAIL assumes a mixture of expert policies $\pi_E = \{\pi_E^0, \pi_E^1, \dots\}$ and specifies a generative process of expert trajectories τ_E based on GAIL, as:

$$s_0 \sim \rho_0, c \sim p(c), \pi \sim p(\pi|c), a_t \sim \pi(a_t|s_t, c), s_{t+1} \sim (s_{t+1}|a_t, s_t)$$

where the policies $\pi(a|s, c)$ are also conditioned to the discrete latent variable c.

Li et al 2017, "InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The InfoGAIL case**

InfoGAIL seeks to **maximize the mutual information between latent codes and trajectories**, denoted $I(c;\tau)$, introducing the variational lower bound

$$L_I(\pi, q) = \mathbb{E}_{c\sim p(c), a\sim\pi(\cdot|s,c)}[log q(c|\tau)] + H(c) \leq I(c;\tau)$$

where $q(c|\tau) \approx q(c|s,a)$ is an approximation of the true posterior $P(c|\tau)$.

Li et al 2017, "InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The InfoGAIL case**

InfoGAIL objective:

$$L_I(\pi, q) = \mathbb{E}_{c\sim p(c), a\sim \pi(\cdot|s,c)}[log q(c|\tau)] + H(c) \leq I(c;\tau)$$

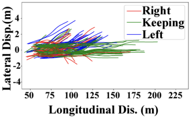
$$min_{\pi_\theta, q} max_D \mathbb{E}_{\pi_\theta}[log(D(s,a))] + \mathbb{E}_{\pi_E}[log(1-D(s,a))] - \lambda_1 \mathcal{H}(\pi_\theta) - \lambda_1 L_I(\pi_\theta, q)$$

Li et al 2017, "InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations"

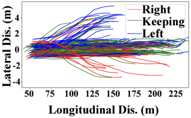
Multimodal Behaviour for continuous actions

How to resolve this problem?

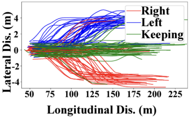
- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The Triple-GAIL case**



BC



GAIL



Triple-GAIL

Fei et al, 2020,"Triple-GAIL: A Multi-Modal Imitation Learning framework with Generative Adversarial Nets"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The Triple-GAIL case**

Triple-GAIL consists of three main components:

- a selector $C_\alpha(c|s,a)$, characterizing $p_{C_\alpha}(c|s,a)$
- a generator $\pi_\theta(a|s,c)$, characterizing $p_{\pi_\theta}(a|s,c)$
- a discriminator $D\psi(s,a,c)$

Seeking for an equilibrium between $p_{C_\alpha}(c|s,a)$ and $p_{\pi_\theta}(a|s,c)$, assuming that $p(s,c)$ and $p(s,a)$ can be obtained from the demonstrations and generated data, respectively.

Adversarial game: The generator and the selector play against the discriminator.

Fei et al, 2020,"Triple-GAIL: A Multi-Modal Imitation Learning framework with Generative Adversarial Nets"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The Triple-GAIL case**

Triple-GAIL consists of three main components:

- a selector $C_\alpha(c|s,a)$, characterizing $p_{C_\alpha}(c|s,a)$
- a generator $\pi_\theta(a|s,c)$, characterizing $p_{\pi_\theta}(a|s,c)$
- a discriminator $D_\psi(s,a,c)$

Triple-GAIL objective:

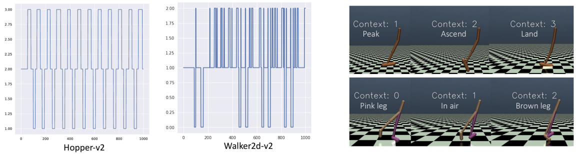
$$
\begin{aligned}
min_{\pi_\theta, C_\alpha} max_{D_\psi} \mathbb{E}_{\pi_\theta}[log(D_\psi(s,a,c))] + \mathbb{E}_{\pi_E}[log(1 - D_\psi(s,a,c))] + \\
(1-\omega)\mathbb{E}_{C_\alpha}[log(D_\psi(s,a,c))] + \\
\lambda_E R_E + \lambda_G R_G - \lambda_H H(\pi_\theta)
\end{aligned}
$$

Fei et al, 2020,"Triple-GAIL: A Multi-Modal Imitation Learning framework with Generative Adversarial Nets"

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The Directed-Info GAIL case**
    Learning intra-trajectory modalities.



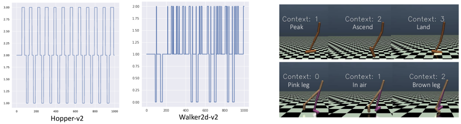
Sharma et al, "Directed-Info GAIL: Learning Hierarchical Policies from Unsegmented Demonstrations using Directed Information"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The Directed-Info GAIL case**
    Learning intra-trajectory modalities.



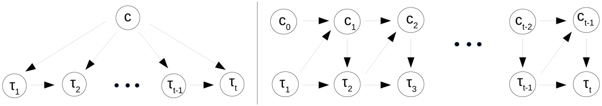


Figure 1: **Left:** Graphical model used in Info-GAIL Li et al. (2017). **Right:** Causal model in this work. The latent code causes the policy to produce a trajectory. The current trajectory, and latent code produce the next latent code

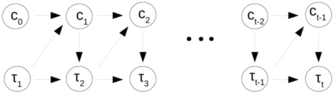
Sharma et al, "Directed-Info GAIL: Learning Hierarchical Policies from Unsegmented Demonstrations using Directed Information"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - **Latent variable models: The Directed-Info GAIL case**
    Learning intra-trajectory modalities.



Directed-Info GAIL objective: Variational lower bound of the directed information:

$$L_1(\pi, q) = \sum_t \mathbb{E}_{c^{1:t} \sim p(c^{1:t}), a^{t-1} \sim \pi(\cdot|s^{t-1}, c^{1:t-1})}[log q(c^t|c^{1:t-1}, \tau^{1:t})] + H(c) \leq I(\tau \rightarrow c)$$

$$min_{\pi_\theta, q} max_D \mathbb{E}_{\pi_\theta}[log(D(s,a))] + \mathbb{E}_{\pi_E}[log(1 - D(s,a))] - \lambda_1 H(\pi_\theta) - \lambda_1 L_1(\pi_\theta, q)$$

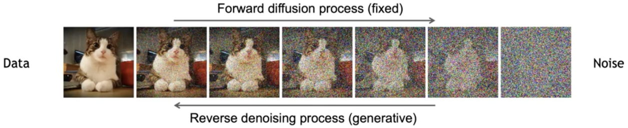
A VAE is pre-trained on the expert trajectories to estimate $p(c^{1:t})$

Sharma et al, "Directed-Info GAIL: Learning Hierarchical Policies from Unsegmented Demonstrations using

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**
    - Forward diffusion process: adds noise to input
    - Reverse denoising process that learns to generate data by denoising



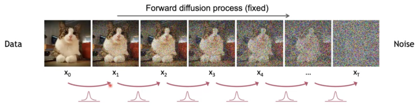
Picture from Jiaming Song et al., "Denoising Diffusion Models: A generative big bang"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**



$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), q(x_N|x_0) \approx \mathcal{N}(x_N; 0, \mathbf{I}))$$

$$q(x_{1:N}|x_0) = \prod_{t=1}^{N} q(x_t|x_{t-1})$$

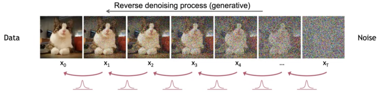
Picture from Jiaming Song et al., "Denoising Diffusion Models: A generative big bang"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**



Goal: Generate data by approx. the denoising model $q(x_{t-1}|x_t)$.

$$p(x_N) = \mathcal{N}(x_N; \mathbf{0}, \mathbf{I})$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t)), \sigma_t^2\mathbf{I}) \rightarrow p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{N} p_\theta(x_{t-1}|x_t)$$

Picture from Jiaming Song et al., "Denoising Diffusion Models: A generative big bang"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**

Adding noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}), \text{ with } q(x_N|x_0) \approx \mathcal{N}(x_N; 0, \mathbf{I}))$$

Denoising:

$$p(x_N) = \mathcal{N}(x_N; \mathbf{O}, \mathbf{I}), \text{ and } p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t)), \sigma_t^2 \mathbf{I})$$

Different choices:

- $x_i = a_t^i$
- $x_i = \tau_i = [(s_0^i, a_0^i), (s_1^i, a_1^i), ..., (s_T^i, a_T^i)]$

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**

Noising Actions: $x_i = a_t^i$

$$a_t^0 = \text{ true action}$$

$$a_t^{i+1} = a_t^i + f(s_t, a_t^i), f(s_t, a_t^i) = noise$$

$$Learned\ model = \hat{f}(s_t, a_t)$$

$$a_t^{i-1} = a_t^i - \hat{f}(s_t, a_t^i),$$

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**

Noising Actions: $x_i = a_t^i$

Represent the policy via the reverse process:

$$\pi_\theta(a|s) = p_\theta(a^{0:N}|s) = \mathcal{N}(a^N; \mathbf{0}, \mathbf{I}) \prod_{i=1}^{N} p_\theta(a^{i-1}|a^i, s)$$

where

$p_\theta(a^{i-1}|a^i, s)$ is modeled as a Gaussian distribution $\mathcal{N}(a^{i-1}; \mu_\theta(a^i, s, i), \sigma_i^2 \mathbf{I})$.

Objective: Train the denoising model by sampling from demonstrated trajectories.

Wang et al 2023 "Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning"

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**

Noising Actions: $x_i = a_t^i$

Diffusion Policy: Visuomotor Policy Learning via Action Diffusion

Cheng Chi[*1], Zhenjia Xu[*1], Siyuan Feng[2], Eric Cousineau[2], Yilun Du[3], Benjamin Burchfiel[2], Russ Tedrake [2,3], Shuran Song[1,4]

Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware

Tony Z. Zhao[1] Vikash Kumar[3] Sergey Levine[2] Chelsea Finn[1]
[1] Stanford University [2] UC Berkeley [3] Meta

Imitation Learning

Multimodal Behaviour for continuous actions

How to resolve this problem?

- More expressive continuous distributions
  - Mixture of Gaussians
  - Latent variable models
  - **Diffusion models**

Noising Trajectories: $x_i = \tau^i$

$$p_\theta(\tau^{i-1}|\tau^i) = \mathcal{N}(\tau^{i-1}|\mu_\theta(\tau_i, i), \sigma_i^2 \mathbf{I})$$

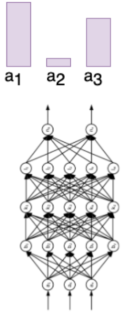
What really happens here?

Janner et al 2022 "Planning with Diffusion for Flexible Behavior Synthesis"

Imitation Learning

Multimodal Behaviour for continuous **and multidimentional** actions

Easy to discretize continuous actions in 1D



but what if we have *n*D?

Imitation Learning

Multimodal Behaviour for continuous **and multidimentional** actions

- Autoregressive discretization

consider multidimensional (nD) actions

$$\mathbf{a}_t = (a_t^1, a_t^2, \dots)$$

We need to learn $\pi_\theta(\mathbf{a}_t|s_t)$:

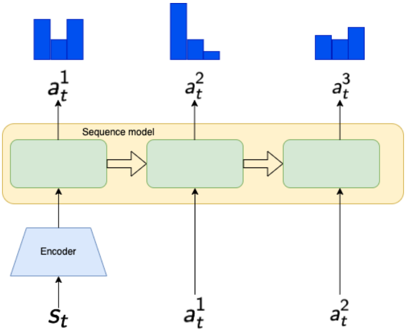
$$\pi_\theta(\mathbf{a}_t|s_t) =$$
$$\pi_\theta(a_t^1, a_t^2, \dots a_t^n|s_t) =$$
$$\pi_\theta(a_t^n|s_t, a_t^0, a_t^1, \dots, a_t^{n-1})\pi_\theta(a_t^{n-1}|s_t, a_t^0, a_t^1, \dots, a_t^{n-2}) \dots \pi_\theta(a_t^1|s_t)$$

Imitation Learning

Multimodal Behaviour for continuous **and multidimentional** actions

- Autoregressive discretization

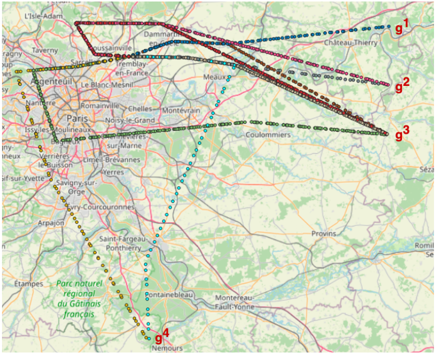
$$\pi_\theta(a_t|s_t) = \pi_\theta(a_t^n|s_t, a_t^0 \ldots a_t^{n-1})\pi_\theta(a_t^{n-1}|s_t, a_t^0 \ldots a_t^{n-2}) \ldots \pi_\theta(a_t^1|s_t)$$



Imitation Learning

Mitigating compounding error via learning many tasks

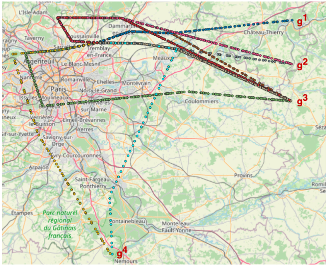
- Already discussed: making mistakes to learn a robust policy
- Learn reaching different goal states



Imitation Learning

Mitigating compounding error via learning many tasks

- Already discussed: making mistakes to learn a robust policy
- **Learn reaching different goal states: Goal Conditioned Behavioural Cloning**



Learn: $\pi_\theta(a|s,g)$, where g is a goal state
Maximizing $log\pi_\theta(a^i_t|s^i_t, g^i = s^i_T)$
Given demos $\{s^i_1, a^i_1, s^i_2, a^i_2, \ldots s^i_{T-1}, a^i_{T-1}, s^i_T\}$ for reaching **goal states** $g^i = s^i_T$

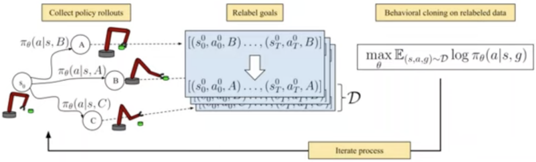
Imitation Learning

Learn reaching different goal states via behavioural cloning

Going **beyond** just imitation?

Learning to Reach Goals via Iterated Supervised Learning

Dibya Ghosh* UC Berkeley
Abhishek Gupta* UC Berkeley
Ashwin Reddy UC Berkeley
Justin Fu UC Berkeley
Coline Devin UC Berkeley
Benjamin Eysenbach Carnegie Mellon University
Sergey Levine UC Berkeley



Imitation Learning

Goal Conditioned Behavioural Cloning

Relay Policy Learning: Solving Long-Horizon Tasks via Imitation and Reinforcement Learning

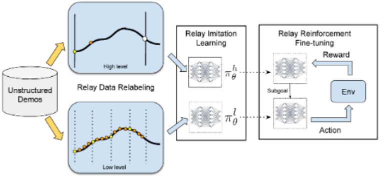
Abhishek Gupta*
Berkeley AI Research, Google
abhigupta@berkeley.edu

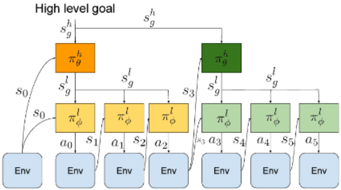
Vikash Kumar
Google
vikashplus@gmail.com

Corey Lynch
Google
coreylynch@google.com

Sergey Levine
Berkeley AI Research, Google
svlevine@eecs.berkeley.edu

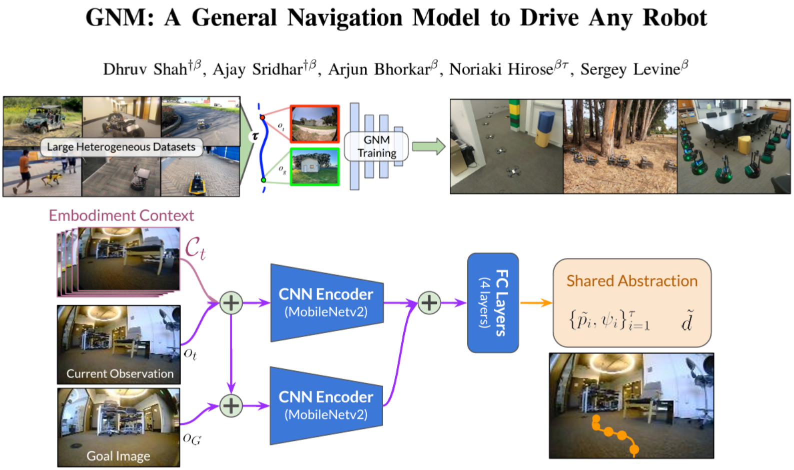
Karol Hausman
Google
karolhausman@google.com





Imitation Learning

Goal Conditioned Behavioural Cloning



The GNM video
The ViNT video