How to Solve Contextual Goal-Oriented Problems with Offline Datasets?

Ying Fan[1], Jingling Li[2, 3], Adith Swaminathan[4], Aditya Modi[4], Ching-An Cheng[4]

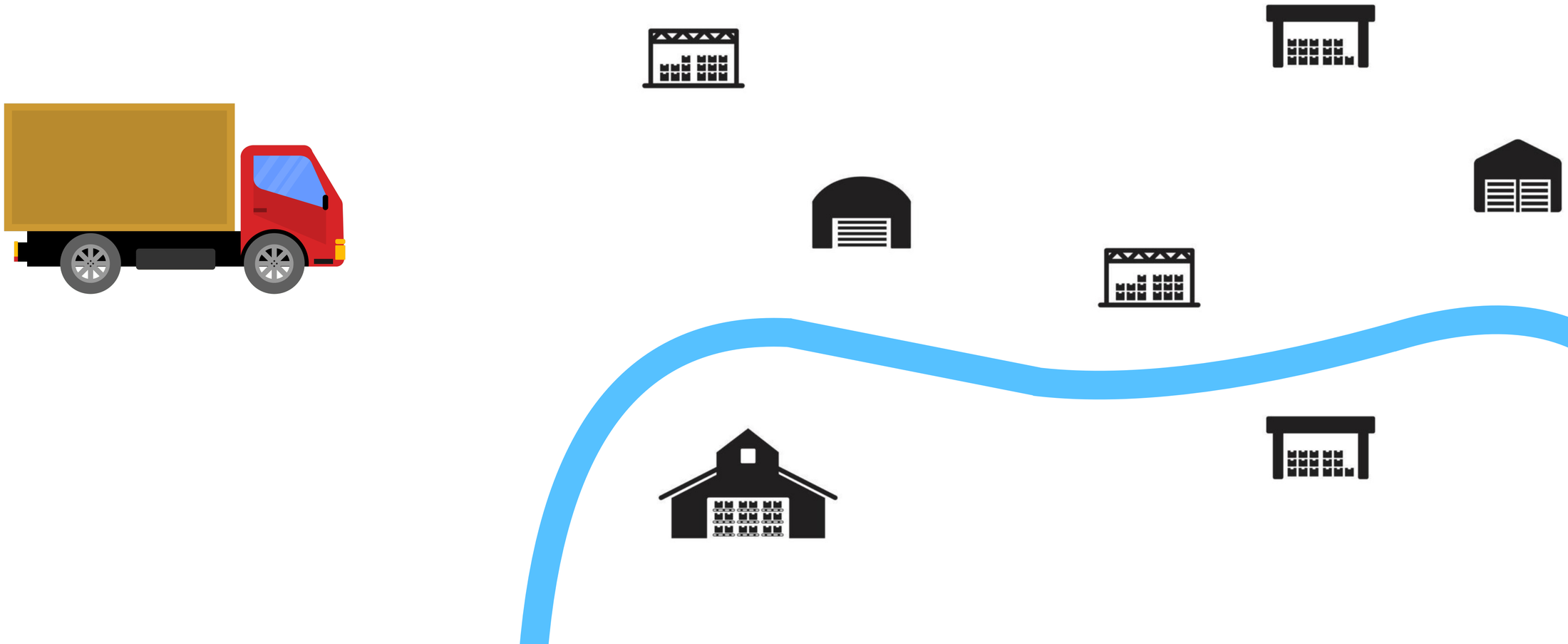
[1] University of Wisconsin–Madison
[2] ByteDance Research
[3] University of Maryland, College Park
[4] Microsoft Research

Work initiated while Ying and Jingling are interns at MSR

Context-based goal-oriented problems (CGO)

Problem Setup

Context: Deliver goods to a warehouse in this area



Context-based goal-oriented problems (CGO)

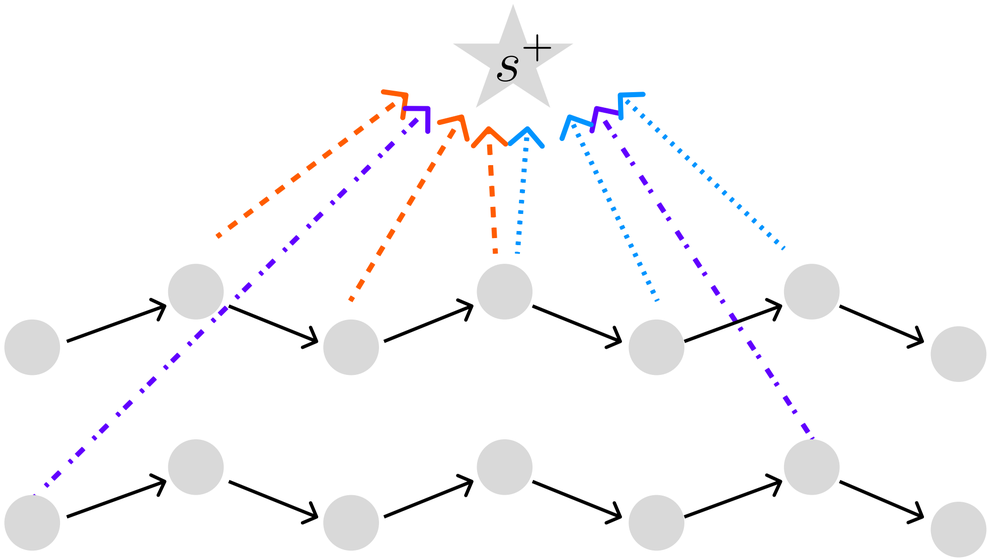
Data Assumption

- **Goal examples**: There is a large amount of **context-goal examples**, where goal examples that are *not necessarily all feasible* (e.g., instruction dataset)
- **Offline experiences** $\{(s, a, s')\}$: The agent has many offline experiences without labels (rewards) – ***dynamics only***
- **Desired outcome**: Given a context, the agent is able to achieve **one of the goals in the corresponding feasible goal set**

Baseline methods & Challenges

- **Goal prediction**
  - We could learn some **goal prediction policy** to predict a goal given context
  - With the dynamics only dataset, we can learn to goal-conditioned policy **(HER)**
  - However, **the predicted goal is not necessarily feasible**!
- **Reward learning**
  - We could form the problem **as missing labels** in the dynamics dataset (given a context)
  - We could **learn a (pessimistic) reward model** with the context-goal dataset (*positive samples only*)
  - **Learning a pessimistic reward model is non-trivial**; also ignores the **goal oriented nature**

Contextual goal-Oriented Data Augmentation (CODA)



: Non-terminal states

: Terminal state

: Real transitions with real actions, unlabeled

: Fictitious transitions given context 2, with fictitious action a^+

: Fictitious transitions given context 1, with fictitious action a^+

: Fictitious transitions given context 3, with fictitious action a^+

- We can **convert** the context-goal dataset to an offline RL dataset:
  - Core idea: given a context, create **"fake" transitions** from the goal examples to a "fake" terminal state with a "fake" action **with reward 1**
- Also, remove all terminal signals in the original transitions, **label with reward 0**, and pair with contexts
- Combine the two, then we naturally have **a fully labeled dataset**

(Contexts could be continuous; do not require exact match to connect true and fake transitions)

Theoretical guarantee

- **Regret equivalence**

Theorem 4.1 (Informal)**.** *The regret of a policy extended to the augmented MDP is equal to the regret of the policy in the original MDP, and any policy defined in the augmented MDP can be converted into that in the original MDP without increasing the regret. Thus, solving the augmented MDP can yield correspondingly optimal policies for the original problem.*

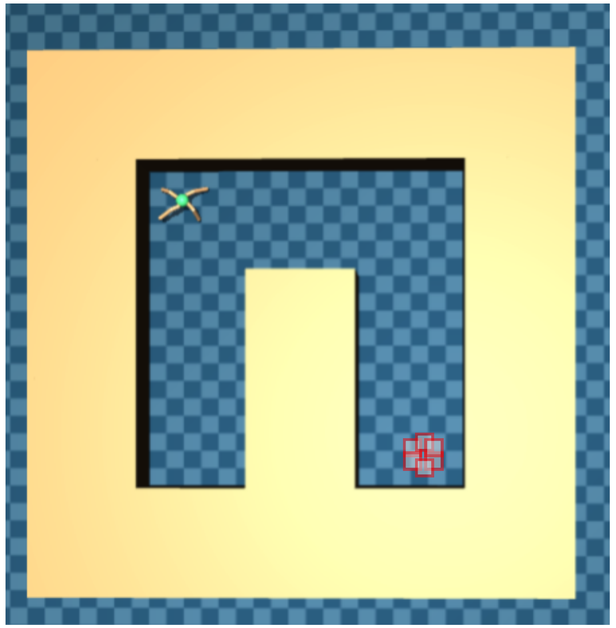
- **Performance guarantee**

Theorem 5.4. *Let $\pi^\dagger$ denote the learned policy of CODA + PSPI with datasets D_{dyn} and D_{goal}, using value function classes $\mathcal{F} = \{\mathcal{X} \times \mathcal{A} \to [0,1]\}$ and $\mathcal{G} = \{\mathcal{X} \to [0,1]\}$. Under Assumption 5.1, 5.2 and 5.3, with probability $1-\delta$, it holds, for any $\pi \in \Pi$,*

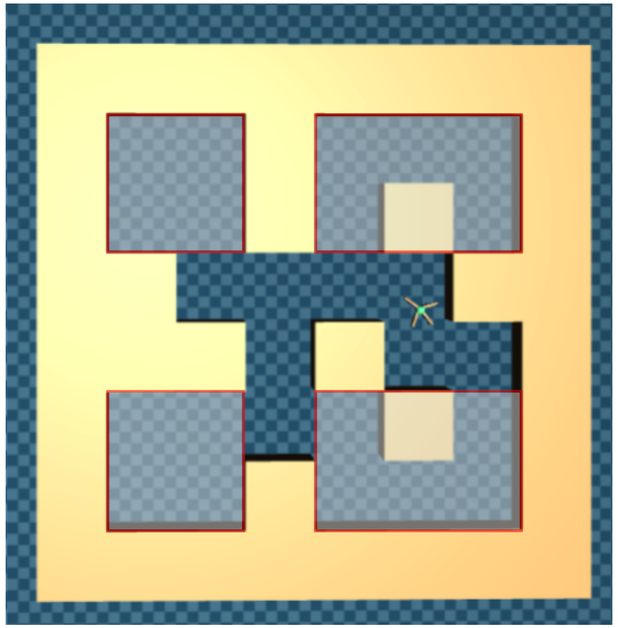
$$J(\pi) - J(\pi^\dagger) \lesssim \mathfrak{C}_{dyn}(\pi)\left(\sqrt{\frac{\log(|\mathcal{F}||\mathcal{G}||\Pi|/\delta)}{|D_{dyn}|}} + \sqrt{\frac{\log(|\mathcal{F}||\mathcal{G}||\Pi|/\delta)}{|D_{goal}|}}\right) + \mathfrak{C}_{goal}(\pi)\sqrt{\frac{\log(|\mathcal{G}|/\delta)}{|D_{goal}|}}$$

where $\mathfrak{C}_{dyn}(\pi)$ and $\mathfrak{C}_{goal}(\pi)$ are concentrability coefficients[4].

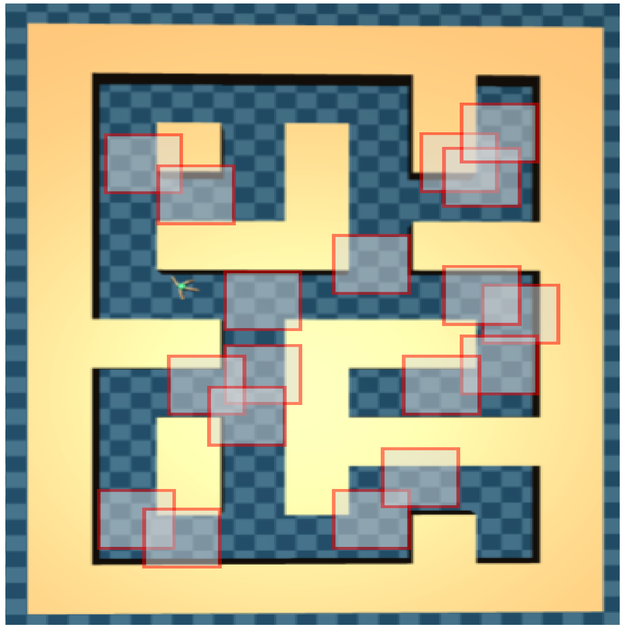
Experiment Settings



Context and goals
are very similar



Different contexts (finite)
maps to distinct goals



The mapping between
contexts (continuous, infinite)
and goals have no constraints

Empirical Results

Table 1: Average success rate (%) in AntMaze-v2 from all environments.

Env/Method	CODA (Ours)	PDS	Goal Prediction	RP	UDS+RP	Oracle Reward
umaze	**94.8±1.3**	93.0±1.3	46.4±6.0	50.5±2.1	54.3±6.3	94.4±0.61
umaze diverse	**72.8±7.7**	50.6±7.8	42.8±4.4	**72.8±2.6**	71.5±4.3	76.8±5.44
medium play	**75.8±1.9**	66.8±4.9	43.8±4.7	0.5±0.3	0.3±0.3	80.6±1.56
medium diverse	**84.5±5.2**	22.8±2.4	28.6±3.9	0.5±0.5	0.8±0.5	72.4±4.26
large play	**60.0±7.6**	39.6±4.9	13.0±4.0	0±0	0±0	41.2±3.58
large diverse	**36.8±6.9**	30.0±5.3	12.6±2.7	0±0	0±0	34.2±2.59
average	**70.8**	50.5	31.2	20.7	21.2	66.6

Table 2: Average scores from Four Rooms with perturbation. The score for each run is the average success rate (%) of the other three rooms.

Env/Method	CODA (Ours)	PDS	Goal Prediction	Oracle Reward
medium-play	**78.7±0.9**	46.0±4.47	59.3±2.6	77.7±2.0
medium-diverse	**83.6±1.9**	51.3±3.6	66.7±2.4	87.4±1.2
large-play	**65.5±2.5**	13.9±2.4	41.4±3.6	67.2±2.7
large-diverse	**72.2±2.9**	11.1±3.8	42.0±3.0	69.6±3.1
average	**75.0**	30.6	52.4	75.5

Table 3: Average scores from Random Cells. The score for each run is the average success rate (%) of random test contexts from the same training distribution.

Env/Method	CODA (Ours)	PDS	Goal Prediction	Oracle Reward
medium-play	**76.8±6.1**	52.0±8.8	66.7±7.2	71.9±0.1
medium-diverse	**78.2±6.5**	60.9±11.3	69.7±8.7	79.3±6.1
large-play	**57.6±12.4**	50.6±6.4	42.4±8.2	49.4±9.3
large-diverse	54.7±8.8	**58.3±9.2**	44.2±8.1	58.2±3.4
average	**66.8**	55.5	55.8	64.7

Thank you!