In-Context Multi-Armed Bandits via Supervised Pretraining

Jiaxin Ye

UCSD Math Department

November 17, 2023

Outline

1. Introduction
2. Model and Theory
3. Algorithm
4. Results and Analysis

Introduction

Context: The Original Paper[1]

- Focus on **transformer-based models** for supervised learning.
- Exploration of **in-context learning** in Reinforcement Learning.
- Introduction of the **Decision-Pretrained Transformer (DPT)**.
- Potential in **sequential decision-making** and **regret guarantees**.

Our Research: Divergence and Novel Contributions

- Builds on the above work.
- **Deviates** from the assumption of sampling from the optimal policy.
- Introduces **imitation learning loss** and **reward reweighting**.
- Focuses on **practical applicability** in real-world scenarios.
- Aims for **minimal performance loss** in offline datasets.

[1] Lee et al., Supervised Pretraining Can Learn In-Context Reinforcement Learning,

Related Concepts

In-context Learning

- **Transformative Approach:** In-context learning enables models to generalize from limited examples by extracting knowledge from the context.
- **Adaptation and Generalization:** Models adapt to various tasks using suitable contextual prompts without parameter updates.
- **Application in Decision-Making:** Utilize state-action-reward tuples to understand interactions with unknown environments.
- **Understanding Dynamics:** Leverage these interactions to comprehend dynamics and identify actions for favorable outcomes.

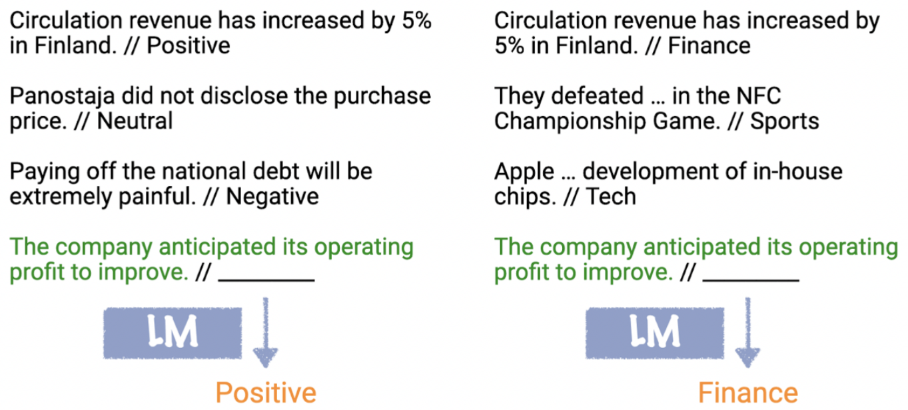
Reward Reweighting

- **Influence on Learning Dynamics:** Reward weighting alters learning, promoting behaviors associated with higher rewards.

More In-context Learning

- Origin: Popularized in the original GPT-3 paper.
- Process:
  1. Give the LM a prompt that consists of a list of input-output pairs that demonstrate a task.
  2. At the end of the prompt, append a test input
  3. Predict the next tokens conditioning on the prompt.
- Model's Task:
  - Understand the input distribution.
  - Recognize the output distribution.
  - Determine the input-output mapping.
  - Comprehend the formatting of the input and output.

Illustration



Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

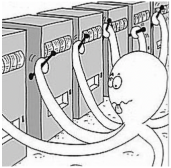
Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // ________

For example, the LM uses the training examples to internally figure out that the task is either sentiment analysis (left) or topic classification (right) and apply the same mapping to the test input.

Decision model: multi-armed bandit

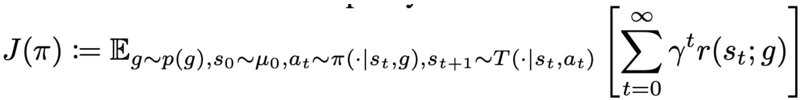
- Specified by a tuple $\xi = \langle A, R \rangle$, where A is the action space and $R : A \to \Delta(\mathbb{R})$ is the reward function.
- State space is trivial (a single state), with no state transitions.
- The process:
  1. At each step t, the agent selects an action a_t from A.
  2. A reward $r_t \sim R(\cdot|a) = N(\mu_a, \sigma^2)$ s.t. $\mu_a = \text{unif}\,[0,1]$ and $\sigma = 0.3$.
- Policy π maps from the single state to a probability distribution over actions, determining which arm to pull.
- The optimal policy π^* maximizes the expected total reward $V(\pi^*) = \max_\pi V(\pi) = \max_\pi \mathbb{E}\left[\sum_t r_t\right]$.
- Learn to decide which arm to pull to maximize cumulative reward with limited knowledge about true reward distributions.



Problem Formulation (Theory Borrowed)[2]

Goal-Conditioned Reinforcement Learning:

- Goal space G defined with a state-to-goal mapping ϕ.
- In our case, to fit into in-context learning, the goal is defined as the trajectory history.
- Reward function $r(s, g)$ and policy $\pi(a|s, g)$ depend on the goal g.
- Objective is to maximize the discounted return: $J(\pi)$ where γ is the discount factor. Usually, $\gamma = 0.9$



$$J(\pi) := \mathbb{E}_{g\sim p(g), s_0\sim\mu_0, a_t\sim\pi(\cdot|s_t,g), s_{t+1}\sim T(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t;g)\right]$$

[2] Ma et al., Offline Goal-Conditioned Reinforcement Learning via f-Advantage Regression

Goal-Conditioned State-Action Occupancy

Goal-Conditioned Occupancy Distribution:

- Defined as $d^{\pi}(s, a; g)$

$$d^{\pi}(s, a; g) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^{t} \Pr(s_t = s, a_t = a \mid s_0 \sim \mu_0, a_t \sim \pi(s_t; g), s_{t+1} \sim T(s_t, a_t))$$

- Captures the relative frequency of state-action visitations conditioned on a goal.
- Subject to the Bellman flow constraint

$$\sum_{a} d(s, a; g) = (1-\gamma)\mu_0(s) + \gamma \sum_{\tilde{s}, \tilde{a}} T(s \mid \tilde{s}, \tilde{a}) d(\tilde{s}, \tilde{a}; g), \qquad \forall s \in S, g \in G$$

Algorithm Introduction

Goal-Conditioned f-Advantage Regression (GoFAR):
Maximizing the discounted return = state-occupancy matching objective

Theorem

Given any function $r(s;g)$, define the target distribution $p(s;g) = \frac{e^{r(s;g)}}{Z(g)}$, where $Z(g) := \int e^{r(s;g)} ds$ is the normalizing constant. Then:

$$-D_{KL}(d^{\pi}(s;g)||p(s;g)) + C = (1-\gamma)J(\pi) + \mathcal{H}(d^{\pi}(s;g))$$

where $C := \mathbb{E}_{g\sim p(g)}[\log Z(g)]$ and $\mathcal{H}(d^{\pi}(s;g)) = \mathbb{E}_{d^{\pi}(s;g)}[\log d^{\pi}(s;g)]$

Definition

The f-divergence of p and q is: (for KL-divergence, $f(x) = x \log x$)

$$D_f(p||q) = \mathbb{E}_{x\sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right]$$

Proof of Theorem

Proof. We have that

$$
\begin{aligned}
&(1-\gamma)J(\pi)\\
=&\mathbb{E}_{g\sim p(g)}\mathbb{E}_{s\sim d^{\pi}(s;g)}\left[r(s;g)\right]\\
=&\mathbb{E}_{g\sim p(g)}\mathbb{E}_{s\sim d^{\pi}(s;g)}\left[\log e^{r(s;g)}\right]\\
=&\mathbb{E}_{g\sim p(g)}\mathbb{E}_{s\sim d^{\pi}(s;g)}\left[\log \frac{e^{r(s;g)}Z(g)}{Z(g)}\right]\\
=&\mathbb{E}_{g\sim p(g)}\mathbb{E}_{s\sim d^{\pi}(s;g)}\left[\log \frac{e^{r(s;g)}}{Z(g)}\right]+\mathbb{E}_{g\sim p(g)}[\log Z(g)]\\
=&\mathbb{E}_{g\sim p(g)}\mathbb{E}_{s\sim d^{\pi}(s;g)}\left[\log \frac{e^{r(s;g)}}{Z(g)}\cdot\frac{d^{\pi}(s;g)}{d^{\pi}(s;g)}\right]+C\\
=&\mathbb{E}_{g\sim p(g)}\mathbb{E}_{s\sim d^{\pi}(s;g)}\left[\log \frac{p(s;g)}{d^{\pi}(s;g)}\right]+\mathbb{E}_{g\sim p(g)}\mathbb{E}_{d^{\pi}(s;g)}[\log d^{\pi}(s;g)]+C\\
=&\mathbb{E}_{g\sim p(g)}\left[-\mathrm{D}_{\mathrm{KL}}(d^{\pi}(s;g)\|p(s;g))-\mathcal{H}(d^{\pi}(s;g))\right]+C
\end{aligned}
$$

Rearranging the inequality gives the desired result.

Theorem

Theorem

For any f-divergence that upper bounds the KL-divergence,

$$-D_{KL}(d^{\pi}(s;g)||p(s;g)) \geq$$
$$\mathbb{E}_{(s,g)\sim d^{\pi}(s,g)}\left[\log\frac{p(s;g)}{d^{O}(s;g)}\right] - D_f(d^{\pi}(s,a;g)||d^{O}(s,a;g))$$

- $R(s;g) = \log\frac{p(s;g)}{d^{O}(s;g)}$: reward that encourages visiting states that occur more often in the "expert" state distribution $p(s;g)$ than in the offline dataset
- Utilizes the offline dataset $d^{O}(s;g)$, suitable for offline learning

Proof of Theorem

Lemma B.1. *For any pair of valid occupancy distributions d_1 and d_2, we have*

$$\mathrm{D}_{\mathrm{KL}}(d_1(s;g)\|d_2(s;g) \leq \mathrm{D}_{\mathrm{KL}}(d_1(s,a;g)\|d_2(s,a;g))$$

Proof. This lemma hinges on proving the following lemma first.

Lemma B.2.

$$\mathrm{D}_{\mathrm{KL}}\left(d_1(s,a,s';g)\|d_2(s,a,s';g)\right) = \mathrm{D}_{\mathrm{KL}}\left(d_1(s,a;g)\|d_2(s,a;g)\right)$$

Proof.

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left(d_1(s,a,s';g)\|d_2(s,a,s';g)\right) \\
=&\int_{S\times A\times S\times G} p(g)d_1(s,a,s';g)\log\frac{d_1(s,a;g)\cdot T(s'\mid s,a)}{d_2(s,a;g)\cdot T(s'\mid s,a)}ds'dadsdg \\
=&\int_{S\times A\times S\times G} p(g)d_1(s,a,s';g)\log\frac{d_1(s,a;g)}{d_2(s,a;g)}ds'dadsdg \\
=&\int_{S\times A\times G} p(g)d_1(s,a;g)\log\frac{d_1(s,a;g)}{d_2(s,a;g)}dadsdg \\
=&\mathrm{D}_{\mathrm{KL}}\left(d_1(s,a;g)\|d_2(s,a;g)\right)
\end{aligned}$$

Proof of Theorem

Using this result, we can prove Lemma B.1:

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left(d_1(s,a;g)\|d_2(s,a;g)\right)\\
=&\mathrm{D}_{\mathrm{KL}}\left(d_1(s,a,s';g)\|d_2(s,a,s';g)\right)\\
=&\int_{S\times A\times S\times G} p(g)d_1(s,a,s';g)\log\frac{d_1(s,a;g)\cdot T(s'\mid s,a)}{d_2(s,a;g)\cdot T(s'\mid s,a)}ds'dadsdg\\
=&\int_{S\times A\times S\times G} p(g)d_1(s;g)\pi_1(a\mid s,g)T(s'\mid s,a)\log\frac{d_1(s,a;g)\cdot T(s'\mid s,a)}{d_2(s,a;g)\cdot T(s'\mid s,a)}ds'dadsdg\\
=&\int p(g)d_1(s;g)\pi_1(a\mid s,g)T(s'\mid s,a)\log\frac{d_1(s;g)}{d_2(s;g)}ds'dadsdg\\
+&\int p(g)d_1(s;g)\pi_1(a\mid s,g)T(s'\mid s,a)\log\frac{\pi_1(a\mid s,g)T(s'\mid s,a)}{\pi_2(a\mid s,g)T(s'\mid s,a)}ds'dadsdg\\
=&\int p(g)d_1(s;g)\log\frac{d_1(s;g)}{d_2(s;g)}dsdg+\int p(g)d_1(s;g)\pi_1(a\mid s,g)\log\frac{\pi_1(a\mid s,g)}{\pi_2(a\mid s,g)}dadsdg\\
=&\mathrm{D}_{\mathrm{KL}}\left(d_1(s;g)\|d_2(s;g)\right)+\mathrm{D}_{\mathrm{KL}}\left(\pi_1(a\mid s,g)\|\pi_2(a\mid s,g)\right)\\
\geq&\mathrm{D}_{\mathrm{KL}}\left(d_1(s;g)\|d_2(s;g)\right)
\end{aligned}
$$

Proof of Theorem

Now given these technical lemmas, we have

$$
\begin{aligned}
& D_{\mathrm{KL}}\left(d^\pi(s;g) \| p(s;g)\right) \\
= & \int p(g) d^\pi(s;g) \log \frac{d^\pi(s;g)}{p(s;g)} \cdot \frac{d^O(s;g)}{d^O(s;g)} d s d g, \quad \text{we assume that } d^O(s;g)>0 \text{ whenever } p(s;g)>0. \\
= & \int p(g) d^\pi(s;g) \log \frac{d^O(s;g)}{p(s;g)} d s d g+\int p(g) d^\pi(s;g) \log \frac{d^\pi(s;g)}{d^O(s;g)} d s d g \\
\leq & \mathbb{E}_{(s,g) \sim d^\pi(s,g)}\left[\log \frac{d^O(s;g)}{p(s;g)}\right]+\mathrm{D}_{\mathrm{KL}}\left(d^\pi(s,a;g) \| d^O(s,a;g)\right)
\end{aligned}
$$

where the last step follows from Lemma B.1. Then, for any $\mathrm{D}_f \geq \mathrm{D}_{\mathrm{KL}}$, we have that

$$
-D_{\mathrm{KL}}\left(d^\pi(s;g) \| p(s;g)\right) \geq \mathbb{E}_{(s,g) \sim d^\pi(s,g)}\left[\log \frac{p(s;g)}{d^O(s;g)}\right]-\mathrm{D}_f\left(d^\pi(s,a;g) \| d^O(s,a;g)\right) \quad (31)
$$

Then, since $\mathbb{E}_{(s,g) \sim d^\pi(s,g)}\left[\log \frac{1}{d^O(s;g)}\right] \geq 0$, we also obtain the following looser bound:

$$
-D_{\mathrm{KL}}\left(d^\pi(s;g) \| p(s;g)\right) \geq \mathbb{E}_{(s,g) \sim d^\pi(s,g)}[\log p(s;g)]-\mathrm{D}_f\left(d^\pi(s,a;g) \| d^O(s,a;g)\right) \quad (32)
$$

Optimization Problem

Recall that

$$-D_{KL}(d^{\pi}(s;g)||p(s;g)) \geq$$

$$\mathsf{E}_{(s,g)\sim d^{\pi}(s,g)}\left[\log \frac{p(s;g)}{d^{O}(s;g)}\right] - D_f(d^{\pi}(s,a;g)||d^{O}(s,a;g))$$

Because $p(s;g) \propto e^{r(s;g)}$, $R(s;g) = \log \frac{p(s;g)}{d^{O}(s;g)} \propto r(s;g)$.
Thus, the optimization problem becomes (with the Bellman constraint),

$$\begin{aligned} & \max_{d(s,a;g)\geq 0} \quad \mathbb{E}_{(s,g)\sim d(s,g)}\left[r(s;g)\right] - \mathrm{D}_f(d(s,a;g)\|d^{O}(s,a;g)) \\ \text{(P)} \qquad & \text{s.t.} \quad \sum_a d(s,a;g) = (1-\gamma)\mu_0(s) + \gamma \sum_{\tilde{s},\tilde{a}} T(s \mid \tilde{s},\tilde{a}) d(\tilde{s},\tilde{a};g), \forall s \in S, g \in G \end{aligned}$$

which still requires sampling from $d(s;g)$, not ideal for offline setting. Can reduce to an unconstrained optimization problem to retrieve the value function.

Dual Problem

Theorem

The dual problem to the above theorem is

$$\min_{V(s;g)\geq 0}(1-\gamma)\mathbb{E}_{(s,g)\sim\mu_0,p(g)}[V(s;g)]+$$

$$E_{(s,a,g)\sim d^O}[f_*(R(s;g)+\gamma TV(s,a;g)-V(s;g))]$$

where f_* denotes the convex conjugate function of f, $V(s;g)$ is the Lagrangian vector, and $TV(s,a;g)=\mathbb{E}_{s'\sim T(|s,a)}[V(s';g)]$

Observation: neither expectation depends on samples from d, thus can be estimated entirely using offline data, making it suitable for offline GCRL.

Once obtained the optimal V^*, learn the policy via the following supervised regression update:

$$\max_{\pi}\mathbb{E}_{g\sim p(g)}\mathbb{E}_{(s,a)\sim d^O(s,a;g)}\left[\left(f'_\star(R(s;g)+\gamma\mathcal{T}V^*(s,a;g)-V^*(s;g)\right)\log\pi(a\mid s,g)\right]$$

Proof of Theorem

Proof. We begin by writing the Lagrangian dual of the primal problem:

$$\min_{V(s;g)\geq 0} \max_{d(s,a;g)\geq 0} \mathbb{E}_{(s,g)\sim d(s,g)} \left[\log\left(r(s;g)\right)\right] - \mathrm{D}_f(d(s,a;g) \| d^O(s,a;g)) \\ + \sum_{s,g} p(g)V(s;g)\left((1-\gamma)\mu_0(s) + \gamma \sum_{\tilde{s},\tilde{a}} T(s \mid \tilde{s},\tilde{a}) d(\tilde{s},\tilde{a};g) - \sum_a d(s,a;g)\right) \tag{35}$$

where $p(g)V(s;g)$ is the Lagrangian vector. Then, we note that

$$\sum_{s,g} V(s;g) \sum_{\tilde{s},\tilde{a}} T(s \mid \tilde{s},\tilde{a}) d(\tilde{s},\tilde{a};g) = \sum_{\tilde{s},\tilde{a},g} d(\tilde{s},\tilde{a};g) \sum_s T(s \mid \tilde{s},\tilde{a}) V(s;g) = \sum_{s,a,g} d(s,a;g)\mathcal{T}V(s,a;g) \tag{36}$$

Using this, we can rewrite (35) as

$$\min_{V(s;g)\geq 0} \max_{d(s,a;g)\geq 0} (1-\gamma)\mathbb{E}_{(s,g)\sim(\mu_0,p(g))}[V(s;g)] + \mathbb{E}_{(s,a,g)\sim d}\left[(r(s;g) + \gamma\mathcal{T}V(s,a;g) - V(s;g))\right] \\ - \mathrm{D}_f(d(s,a;g) \| d^O(s,a;g)) \tag{37}$$

Proof of Theorem

And finally,

$$\min_{V(s,g)\geq 0}(1-\gamma)\mathbb{E}_{(s,g)\sim(\mu_0,p(g))}[V(s;g)] + \max_{d(s,a;g)\geq 0}\mathbb{E}_{(s,a,g)\sim d}\left[(r(s,g)+\gamma\mathcal{T}V(s,a;g)-V(s;g))\right] - \mathrm{D}_f(d(s,a;g)\|d^O(s,a;g)) \tag{38}$$

Now, we make the key observation that the inner maximization problem in (38) is in fact the Fenchel conjugate of $\mathrm{D}_f(d(s,a,g)\|d^O(s,a,g))$ at $r(s,g)+\gamma\mathcal{T}V(s,a,g)-V(s,g)$. Therefore, we can reduce (38) to an unconstrained minimization problem over the dual variables

$$\min_{V(s,g)\geq 0}(1-\gamma)\mathbb{E}_{(s,g)\sim\mu_0,p(g)}[V(s;g)] + \mathbb{E}_{(s,a,g)\sim d^O}\left[f_\star\left(r(s,g)+\gamma\mathcal{T}V(s,a;g)-V(s;g))\right]\right., \tag{39}$$

and consequently, we can relate the dual-optimal V^* to the primal-optimal d^* using Fenchel duality (see Appendix A:

$$d^*(s,a;g) = d^O(s,a;g)f'_\star\left(r(s,g)+\gamma\mathcal{T}V^*(s,a,g)-V^*(s,g)\right), \forall s\in S, a\in A, g\in G, \tag{40}$$

as desired. □

Adapted to the Bandit Case

$$\max_{\pi} \mathbb{E}_{g\sim p(g)} \mathbb{E}_{(s,a)\sim d^{O}(s,a;g)} \left[(f'_{\star}(R(s;g) + \gamma \mathcal{T} V^{*}(s,a;g) - V^{*}(s;g)) \log \pi(a \mid s,g) \right]$$

Since we're considering the bandit case where there is only one state, there is no discount factor γ needed. Also, since there are no future states—only immediate rewards—the "value" of taking an action is typically just the expected immediate reward of that action. Thus, training for a optimal policy is reduced to:

$$\max_{\pi} \mathbb{E}_{g\sim p(g)} \mathbb{E}_{(a)\sim d^{O}(a;g)} [(f'_{*}(R(a;g)) \cdot \log \pi(a|g)]$$

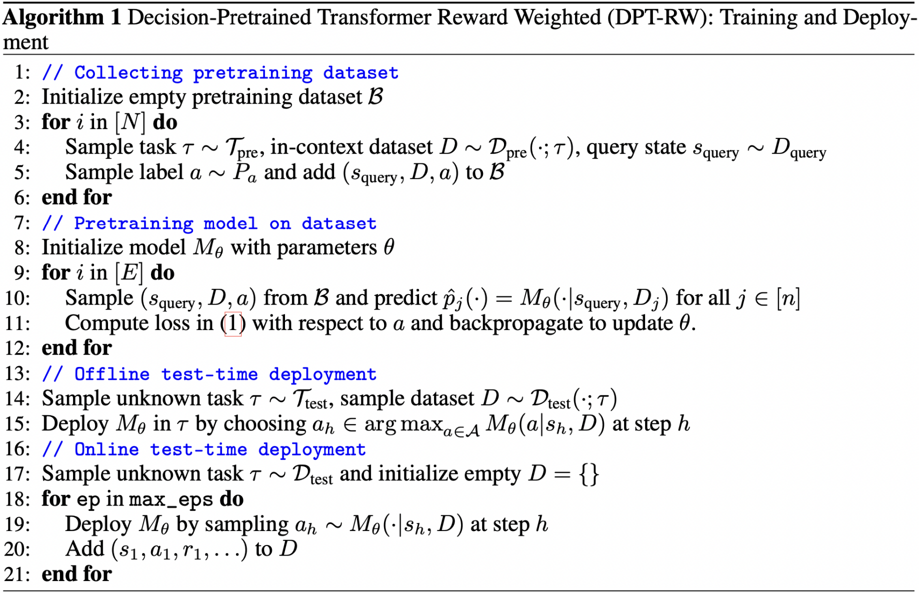
For KL-divergence, $f(x) = x \log x$, $D_{*,f}(y) = \log \mathbb{E}_{x\sim q}[\exp y(x)]$, so

$$\max_{\pi} \mathbb{E}_{g\sim p(g)} \mathbb{E}_{(a)\sim d^{O}(a;g)} [\exp(R(a;g) \cdot \log \pi(a|g)]$$

The loss function is the following for the bandit case: (reward reweighting)

$$L(\pi) = -\sum_{j\in[n]} [\exp(R_j(a;g)) \cdot \log \pi(a|g)]$$

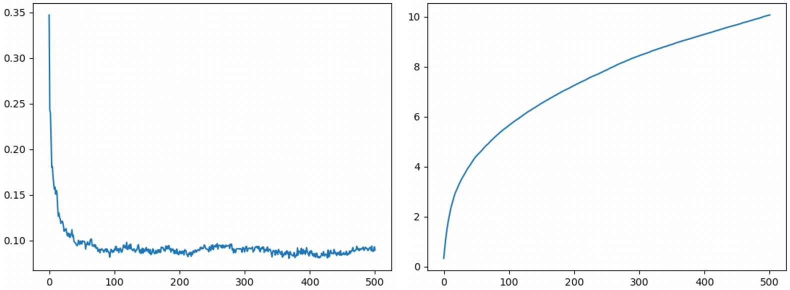
Algorithm



Algorithm 1 Decision-Pretrained Transformer Reward Weighted (DPT-RW): Training and Deployment

```
// Collecting pretraining dataset
Initialize empty pretraining dataset $\mathcal{B}$
for $i$ in $[N]$ do
    Sample task $\tau \sim \mathcal{T}_{\text{pre}}$, in-context dataset $D \sim \mathcal{D}_{\text{pre}}(\cdot;\tau)$, query state $s_{\text{query}} \sim D_{\text{query}}$
    Sample label $a \sim P_a$ and add $(s_{\text{query}}, D, a)$ to $\mathcal{B}$
end for
// Pretraining model on dataset
Initialize model $M_\theta$ with parameters $\theta$
for $i$ in $[E]$ do
    Sample $(s_{\text{query}}, D, a)$ from $\mathcal{B}$ and predict $\hat{p}_j(\cdot) = M_\theta(\cdot|s_{\text{query}}, D_j)$ for all $j \in [n]$
    Compute loss in (1) with respect to $a$ and backpropagate to update $\theta$.
end for
// Offline test-time deployment
Sample unknown task $\tau \sim \mathcal{T}_{\text{test}}$, sample dataset $D \sim \mathcal{D}_{\text{test}}(\cdot;\tau)$
Deploy $M_\theta$ in $\tau$ by choosing $a_h \in \arg\max_{a \in \mathcal{A}} M_\theta(a|s_h, D)$ at step $h$
// Online test-time deployment
Sample unknown task $\tau \sim \mathcal{D}_{\text{test}}$ and initialize empty $D = \{\}$
for ep in max_eps do
    Deploy $M_\theta$ by sampling $a_h \sim M_\theta(\cdot|s_h, D)$ at step $h$
    Add $(s_1, a_1, r_1, \ldots)$ to $D$
end for
```

Test Results



Left (offline): suboptimality $\mu_a^* - \mu_{\hat{a}}$ where $\hat{a}$ is the chosen action; achieving a performance metric of 0.1 after 50 steps
Right (online): cumulative regret $\sum_k \mu_a^* - \mu_{\hat{a}_k}$ where $\hat{a}_k$ is the kth chosen action; exhibiting a logarithmic trend, with final regret just above 10