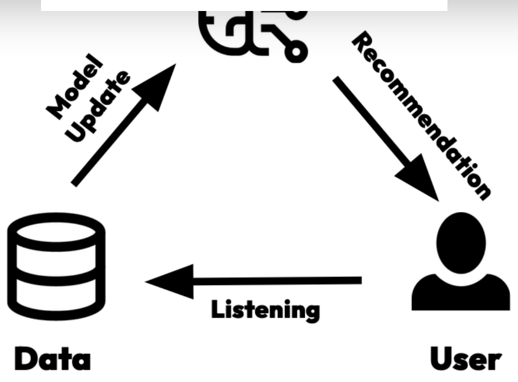
In-context Exploration-Exploitation for Reinforcement Learning

May 07, 2024
Published by Zhenwen Dai, Federico Tomasi, Sina Ghiassian

Machine learning (ML) models are widely used on Spotify to provide users with a daily personalized listening experience. To ensure these systems continually adapt to users' preferences and swiftly adjust to changes in their interests, the ML systems need to rapidly update themselves with incoming user interaction data. The process of updating an ML model is illustrated in the following diagram. Ideally, the shorter the update cycle, the quicker the ML systems can learn about users' preference changes. However, in typical production ML systems, a model update cycle can range from several hours to a few days.



As depicted in the diagram, the update cycle is primarily driven by the ML model's recommendations. A high-quality recommendation not only suggests content that aligns with a user's preferences but also selects content that efficiently gathers information about the user's real-time interest. This enables the ML model to respond quickly to a user's needs. A principled approach to tackle this challenge is known as Bayesian decision making. This method addresses the problem by maintaining a Bayesian belief about a user's interest and selecting content that decreases the uncertainty of this belief while catering to the user's interest. This balancing act is often referred to as the exploration-exploitation (EE) trade-off.

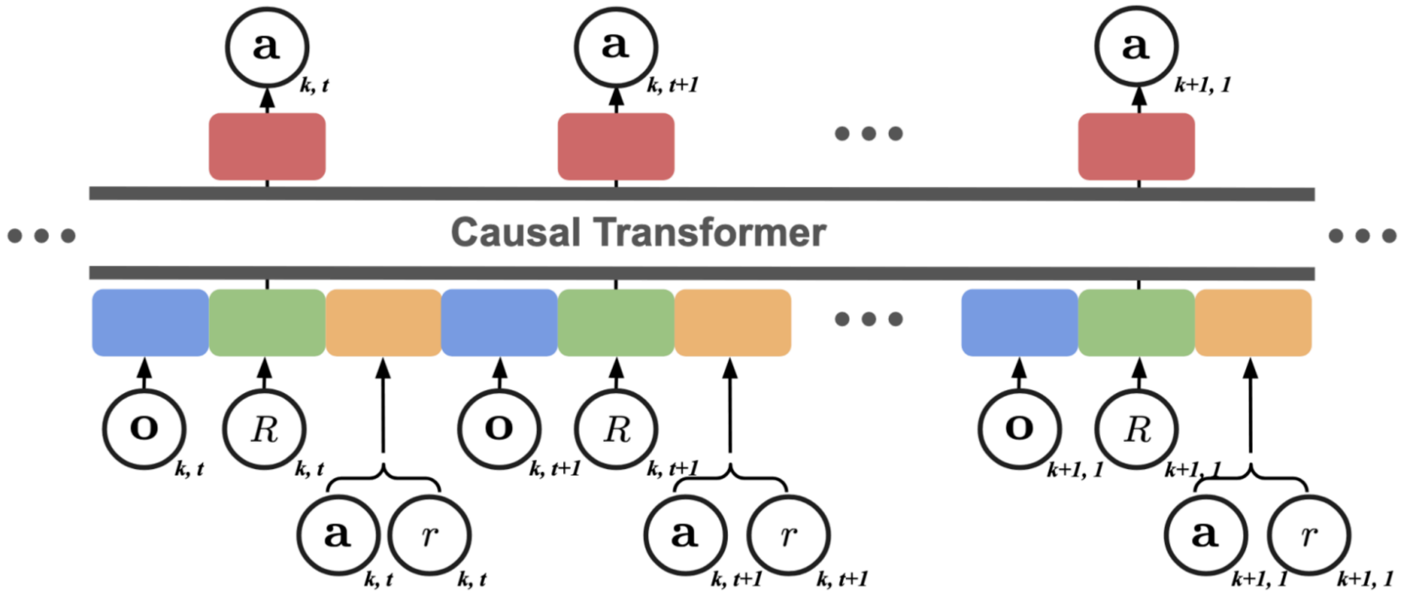
Conventional methods in this field accomplish the Bayesian belief update through model updating, typically implemented via gradient-based optimization. While being a well established theory framework, Bayesian modeling approaches suffer from high computational complexity and various other limitations such as hard to specify good prior distribution.

In this work, we present a groundbreaking approach to model updating known as in-context exploration exploitation (ICEE). ICEE allows us to achieve Bayesian belief updates through neural network inference, made possible by adopting the concept of in-context learning. With this approach, Bayesian decision making can be accomplished through standard supervised learning and neural network inference, enabling real-time adaptation to users' preference. The following sections provide a detailed explanation of ICEE.

Return conditioned In-context Policy Learning

distinct task due to the significant variation in users' content preferences. Each user's listening session can be interpreted as an episode in the episodic RL formulation as the user's intent may differ from session to session, leading to varied activities. These activities can generate statistics used to define the reward for our agent. This setting is commonly referred to as meta RL.

ICEE is designed to address the aforementioned meta RL problem with one additional constraint: it cannot perform parameter updates while learning to solve individual tasks. ICEE tackles this challenge by expanding the framework of return conditioned RL with in-context learning. In this formulation, both policy learning and action prediction are modeled as a sequence prediction problem. For each task, the information of all the episodes are consolidated into a single sequence, as illustrated in the figure below.



In this sequence, a time step in an episode is represented by a triplet: state, return-to-go, and a combination of the action taken and the resulting reward. The return-to-go indicates the agent's future performance. The time steps within an episode are arranged chronologically. There is no specific requirement about the order of episodes. For simplicity, we also arrange the episodes in chronological order. This sequence is then fed into a Transformer decoder architecture with a causal attention mechanism. The Transformer decoder's output, which corresponds to the return-to-go's location, is used to predict the action for the corresponding time step.

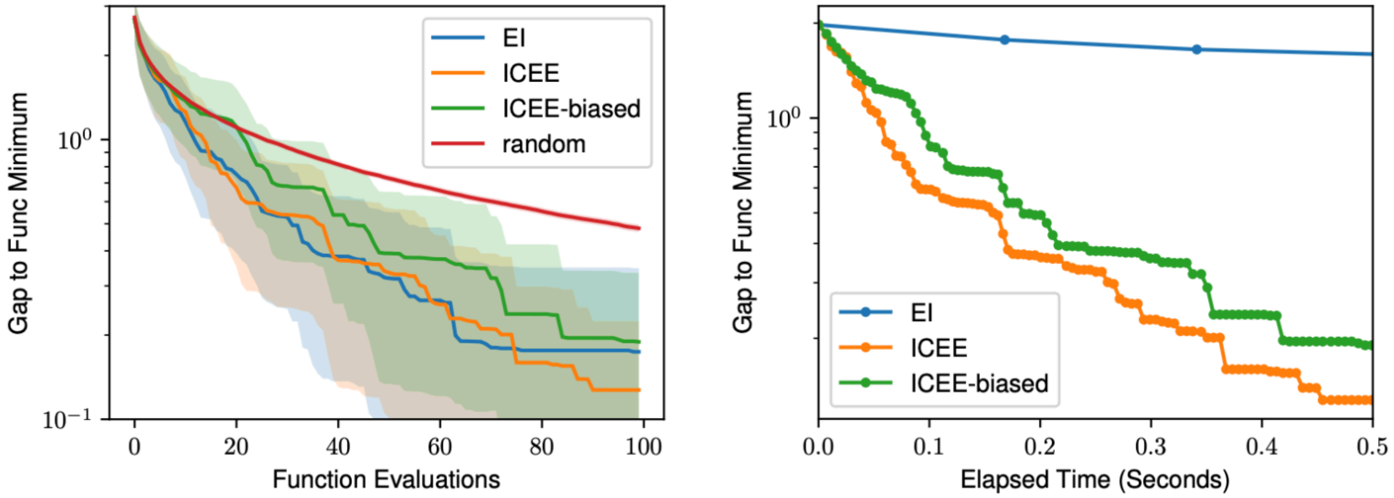
from the model. After applying the action in the environment, we observe the reward and the next state. This new information is added to the input sequence and the next action is sampled. This interaction loop continues until the episode concludes. After completing the first episode, we begin with the initial state of a new episode and repeat the process. Through this method, the model can solve a new task after a few episodes. It's clear from this description that no model parameter updates occur during the task-solving process. All learning occurs by collecting information through actions and using this information to inform future actions.

A key component in the above description is the choice of return-to-go. It enables the model to differentiate between good and bad actions during training and to only select good actions during inference. In ICEE, the return-to-go consists of in-episode return-to-go and cross-episode return-to-go. For in-episode return-to-go, we follow the design of the multi-game decision transformer (MGDT). It defines return-to-go as the cumulative future reward from the current step and includes a model that predicts return at each step. During inference, it samples a value from the return distribution skewed towards good returns and uses this value as the return-to-go. For cross-episode return-to-go, we use a binary value. It's set to one if the current episode's return is better than all previous episodes, and zero otherwise. During inference, the value is set to one while taking actions and adjusted to the true value after the episode ends.

Compared to previous return conditioned RL approaches, the balance between exploration and exploitation is particularly crucial. The model needs to learn to solve a new task with the least number of episodes. To achieve efficient EE, we propose an unbiased training objective for ICEE. This is based on the observation that the action learned from standard return conditioned RL methods is biased towards the data collection policy. By reformulating the training objective, we obtain an unbiased training objective, which corrects the previous training objective with a probability ratio between the uniform action distribution and the data collection distribution.

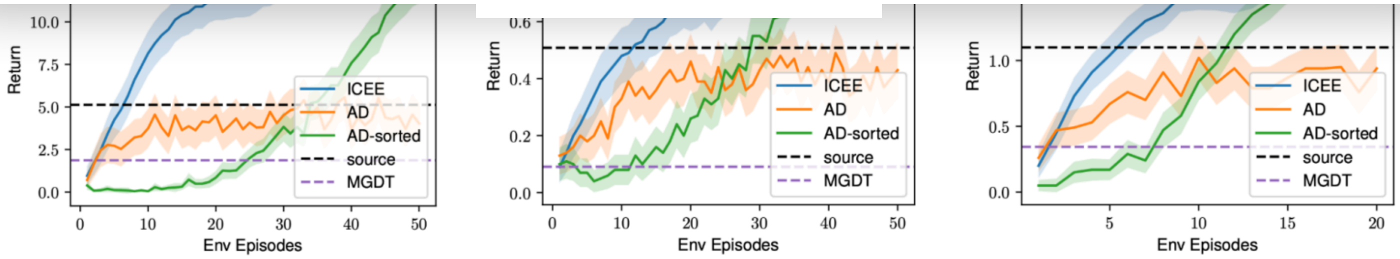
Experiments

evaluations. The Gaussian process (GP) based BO methods have been widely used in various domains like hyper-parameter tuning, drug discovery, aerodynamic optimization. To evaluate the EE performance of ICEE, we apply ICEE to BO and compare it with a GP-based approach using one of most widely used acquisition functions, expected improvement (EI).



The above figures show a comparison of ICEE and EI on a set of 2D benchmark functions. The search efficiency of ICEE is on par with the GP-based BO method with EI. This indicates that ICEE can learn to perform EE through ICL. A clear advantage of ICEE is that the whole search is done through model inference without need of any gradient optimization. In contrast, GP-based BO methods need to fit a GP surrogate function at each step, which results in a significant speed difference.

Grid-world RL. We investigate the in-context policy learning capability of ICEE on sequential RL problems. We focus on the families of environments that cannot be solved through zero-shot generalization of a pre-trained model, so in-context policy learning is necessary for solving the tasks. We use the two grid world environments: dark room and dark key-to-door.



The experiment results, shown in figure above, demonstrate that ICEE is able to solve the sampled games efficiently compared to the baseline methods. The EE capability allows ICEE to search for the missing information efficiently and then acts with confidence once the missing information is found. More details of the above experiments can be found in the paper.

Conclusions

We present an in-context EE algorithm by extending the decision transformer formulation to in-context learning and deriving an unbiased training objective. Through the experiments on BO and discrete RL problems, we demonstrate that: (i) ICEE can perform EE in in-context learning without the need of explicit Bayesian inference; (ii) The performance of ICEE is on par with state-of-the-art BO methods without the need of gradient optimization, which leads to significant speed-up; (iii) New RL tasks can be solved within tens of episodes.

These outcomes show that ICEE could be a promising technology to enhance the responsiveness of personalization experiences by substantially reducing model update cycles.

For more information, please refer to our paper:
In-context Exploration-Exploitation for Reinforcement Learning
Zhenwen Dai, Federico Tomasi, Sina Ghiassian
ICLR 2024

SHARE

CATEGORIES

Machine Learning

Reinforcement Learning

Related articles

PODTILE: Facilitating Podcast Episode Browsing with Auto-generated Chapters

Personalizing Audiobooks and Podcasts with graph-based models

Accelerating Creator Audience Building through Centralized Exploration

Sign up for research updates

By clicking sign up you’ll receive occasional emails from Spotify. You always have the choice to adjust your interest settings or unsubscribe.

Your Email Sign Up