

Policy Gradients for Contextual Recommendations

Feiyang Pan^{1,3}, Qingpeng Cai², Pingzhong Tang², Fuzhen Zhuang^{1,3}, Qing He^{1,3}

¹Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, China.

²IIS, Tsinghua University.

³University of Chinese Academy of Sciences, China.

ABSTRACT

Decision making is a challenging task in online recommender systems. The decision maker often needs to choose a contextual item at each step from a set of candidates. Contextual bandit algorithms have been successfully deployed to such applications, for the trade-off between exploration and exploitation and the state-of-art performance on minimizing online costs. However, the applicability of existing contextual bandit methods is limited by the over-simplified assumptions of the problem, such as assuming a simple form of the reward function or assuming a static environment where the states are not affected by previous actions.

In this work, we put forward *Policy Gradients for Contextual Recommendations (PGCR)* to solve the problem without those unrealistic assumptions. It optimizes over a restricted class of policies where the marginal probability of choosing an item (in expectation of other items) has a simple closed form, and the gradient of the expected return over the policy in this class is in a succinct form. Moreover, PGCR leverages two useful heuristic techniques called Time-Dependent Greed and Actor-Dropout. The former ensures PGCR to be empirically greedy in the limit, and the latter addresses the trade-off between exploration and exploitation by using the policy network with Dropout as a Bayesian approximation.

PGCR can solve the standard contextual bandits as well as its Markov Decision Process generalization. Therefore it can be applied to a wide range of realistic settings of recommendations, such as personalized advertising. We evaluate PGCR on toy datasets as well as a real-world dataset of personalized music recommendations. Experiments show that PGCR enables fast convergence and low regret, and outperforms both classic contextual-bandits and vanilla policy gradient methods.

CCS CONCEPTS

• **Information systems** → **Online advertising**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Recommender System; Policy Gradients; Contextual Bandits; Reinforcement Learning;

Pingzhong Tang (kenshinping@gmail.com) is the corresponding author.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6674-8/19/05.

<https://doi.org/10.1145/3308558.3313616>

ACM Reference Format:

Feiyang Pan^{1,3}, Qingpeng Cai², Pingzhong Tang², Fuzhen Zhuang^{1,3}, Qing He^{1,3}. 2019. Policy Gradients for Contextual Recommendations. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3308558.3313616>

1 INTRODUCTION

Decision making in online recommender systems and advertising systems are challenging because the recommender needs to find the policy that maximizes its revenue by interacting with the world. A typical decision-making problem is to select a featured item from a finite set of candidates, for example, to select an advertisement from a set of ads that relate to the user's query in a search engine, or to recommend a song from the user's playlist in a music streaming service. After making the decision, the recommender will receive a reward together with some state transition. In such settings, each item has a so-called *context* (which includes features and attributes) that carries all the necessary information for making the choice. Since it is often the case that the reward, as well as the state dynamic of choosing each item, are related to its context, the recommender system must try to learn how to make the choice given the contexts of all the candidates.

To solve such contextual recommendation problems, algorithms based on *contextual-bandits* have been successfully deployed in a number of industrial level applications over the past decade, such as personalized recommender systems [17, 27, 28], advertisement personalization [6, 29], and learning-to-rank [22]. Contextual-bandit algorithms are preferred if one needs to minimize the cumulative cost during online-learning because they aim to address the trade-off between exploitation and exploration.

The standard contextual bandit problem can be seen as a repeated game between nature and the player [16]. Nature defines a reward function mapping contexts (a set of features) to real-valued rewards, which is unknown to the player. At each step, nature gives a set of items, each with a context. The player observes the contextual items, selects one, and then receives a reward. The payoff of the player is to minimize the cumulative regret or to maximize the cumulative reward.

The main challenge of solving contextual bandits lies in the trade-off between exploration and exploitation. The most well-known approaches are arguably value-based methods including Upper Confidence Bounds (UCB) [5], Thompson Sampling (TS) [30], and their variants. These value-based methods try to estimate the expected reward of choosing each item, so they are especially effective when the form of reward function is known explicitly. For example, when the expected reward is linear in the context, [1, 9, 17] proposed Lin-UCB, which is applied successfully to the online news

recommendation of Yahoo, and [4, 8, 19] also proposed TS to solve the linear contextual bandits. Similarly, [10] proposed GLM-UCB using generalized linear models, [15, 23] used Gaussian Processes, to model the reward functions. These variants of UCB and TS have been known to achieve sub-linear regrets [1, 2, 4, 5, 8–10, 15, 17, 23]. Similar ideas have also been applied to reinforcement learning algorithm such as the UCRL algorithm with regret bounds [13].

However, the applicability of these approaches in real-world applications is heavily limited, especially for large-scale and high-dimensional problems, due to the following reasons:

- First, these methods tend to over-simplify the form of the reward function, which is unrealistic in real-world cases. For example, for sponsored search advertising via real-time bidding, the reward of showing an ad (cost per click) is often the click-through rate multiplied by the bidding price, so it can be understood as a mixture of binary and linear outcomes. Moreover, the reward can often be a high-order non-linear function of features in the contexts.
- Second, the overall formulations of contextual-bandit problems are sometimes over-simplified comparing to real-world applications. It is often assumed that the reward is determined by the context of the currently chosen item, and the distribution of contexts is independent of the agent’s action. However, it may not be true in real-world recommender systems where the behaviors of users heavily depend on not only the current contexts but the history, i.e., the items that he/she viewed in previous rounds. Also, the set of candidate items can relate to the user’s previous preferences as well. These dependencies are not well exploited in existing models.
- Last but not least, these methods are value-based, so they are meant to find deterministic policies. A subtle change in the value estimation may cause a discontinuous jump in the resulting policy, which makes convergence difficult for these [25]. On the other hand, stochastic policies are sometimes preferred in online recommender systems.

In light of these observations, we propose Policy Gradients for Contextual Recommendations (PGCR), which uses the policy gradient method to solve general contextual recommendations. Our approach model the contextual recommendation problem without unrealistic assumptions or prior knowledge. By optimizing directly over the parameters of stochastic policies, it naturally fits the problems that require randomized actions as well as addresses the trade-off between exploration and exploitation.

Since we design PGCR specifically for contextual recommendations, we would like to specify the performance objective first and see if it is different from the one in standard reinforcement learning. We find that the objective over policies depends on the marginal expected probability of choosing each item (in expectation of other items). So PGCR restricts the search space to a class of policies in which the expected probabilities of choosing an item has a simple closed form and can be estimated efficiently. Therefore, the search space for PGCR is dramatically reduced.

Then, in order to estimate the marginal probability of choosing each item, we extend Experience Replay, the popular technique in

off-policy reinforcement learning [3, 12], to a finer-grained sampling procedure. By doing so, the variance of estimating policy gradients can be smaller than the variance of the vanilla policy gradient algorithm. The resulted algorithm is also computationally efficient by stochastic gradient descent with mini-batch training.

To address the trade-off of exploration and exploitation, our proposed PGCR empirically has the property of Greedy in the Limit with Infinite Exploration (GLIE), which is an essential property for contextual bandits [19]. The property is guaranteed by two useful heuristics named *Time-Dependent Greed* and *Actor-Dropout*. Time-Dependent Greed is to schedule the level of greed to increase over time, so the resulted stochastic policy will explore a lot in the early stage and then gradually converge to a greedy policy. Actor-Dropout is to use dropout on the policy network while training and inferring, thus the feed-forward network outputs policies with randomness. It has been known that such a stochastic feed-forward neural network can be seen as a Bayesian approximation [11], so it can provide with directed exploration for PGCR.

Furthermore, with the mentioned techniques, PGCR can directly apply to contextual recommendations in a Markov Decision Process (MDP) setting, i.e. with states and state transitions. We propose this generalized setting for the reason that the i.i.d. assumption on contexts in the standard contextual bandit setting is unrealistic for real-world applications. On the other hand, we suppose that at each step, the contexts are drawn i.i.d. from a distribution conditional on the current state. Furthermore, when an item is chosen, the immediate reward is determined by both the state and the selected item. The state is then transitioned into the next state. Such a model is tailored for a wide range of important realistic applications such as personalized recommender systems where users’ preferences are regarded as states and items are regarded as items with contexts [20, 26], and e-commerce where the private information (e.g., cost, reputation) of sellers can be viewed as states and different commercial strategies are regarded as contexts [7].

We evaluate PGCR on toy datasets and a real-world dataset of music recommendation. By comparing with several common baselines including Lin-UCB, GLM-UCB, Thompson Sampling, ϵ -greedy, and vanilla policy gradients, it shows that PGCR converges quickly and achieves the lowest cumulative regret and the highest average reward in various standard contextual-bandits settings. Moreover, when state dynamics are included in the real-world recommendation environments, we find that GLM-UCB and TS fail to incorporate information from the states, while PGCR consistently outperforms other baselines.

2 PROBLEM FORMULATION

2.1 One-step Contextual Recommendations

We first introduce the simplified setting of contextual recommendation as a standard contextual-bandits problem. At each step, we have a set of contexts $\mathbf{c} = (c_1, \dots, c_m)^T$ that corresponds to m items, where c_i is the context of the i^{th} item. The contexts c_1, \dots, c_m are independently and identically distributed random variables with outcome space \mathcal{C} . The action is to select an item from the candidates, $a \in \{1, \dots, m\}$.

For the ease of notation, we use \mathbf{c} to denote the concatenation of all m contexts and use c_a to denote the context of the selected

item a . We write the random variable of immediate reward as $R(c_a)$ to note that in this setting it depends only on the chosen context vector c_a . The dependency is not known to the decision-maker. So the target is to learn the dependency and choose the item with the largest expected reward.

A stochastic policy π is a function that maps the observations (the set of contexts \mathbf{c}) to a distribution of actions. Let the random variable $a \sim \pi(\mathbf{c})$ denote the action determined by policy π . The performance of a policy is measured as the expected reward of the chosen item over all possible contexts, i.e.,

$$J(\pi) = \mathbb{E}_{\mathbf{c}} [R(c_a) \mid a \sim \pi(\mathbf{c})], \quad (1)$$

where c_a is short for the context of the chosen action a .

When the policy π is parameterized as π_θ where θ is the trainable parameters, our goal is to find the optimal choice of θ that maximizes the objective $J(\pi_\theta)$.

However, there is an obvious drawback for this simplified setting: in real-world recommendations, it cannot be assumed that the contexts are always drawn i.i.d from some global probability distribution. For example, when recommending items (goods) to a customer given the searching query in an e-commerce platform, we can only select items from a candidate pool related to the query. This non-i.i.d nature motivates us to put forward a more general setting, which involves the states.

2.2 Sequential State-aware Contextual Recommendations

In this part, we introduce the generalized setting as a Markov Decision Process (MDP) with states and state transitions for contextual recommendations, which is referred to as MDP-CR.

At each step t , the decision maker observes its state s_t as well as a set of contexts correlated to that state $\mathbf{c}_t = \{c_{t1}, \dots, c_{tm}\}$. When an action (one of the items) $a_t = \pi(s_t, \mathbf{c}_t)$ is selected, a reward $R(s_t, c_{ta_t})$ is received, and the state is transitioned to the next state by a Markovian state transition probability $s_{t+1} \sim T(s_{t+1} \mid s_t, c_{ta_t})$. Note that the setting in this paper is different from other existing generalized bandits with transitions such as Restless bandits [31].

In this setting, we assume that the contexts are independently distributed conditioning on the state: $c_{ti} \sim g^{s_t}(c)$ for all i , where $g^{s_t}(c)$ is the probability density of contexts given state s_t . For example, if the state is the search query or the attribute vector of a user, we assume that the contexts of items in the candidate pool are drawn from a distribution that reflects the search query or the user preference.

The goal is to find a policy that maximizes the expected cumulative discounted reward, so the objective is

$$J(\pi) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, c_{ta_t}) \mid a_t \sim \pi(s_t, \mathbf{c}_t) \right], \quad (2)$$

where $0 < \gamma < 1$ is a discount factor that balances short and long term rewards, just like in standard reinforcement learning. We also define the action value function

$$Q^\pi(s, \mathbf{c}) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, c_{ta_t}) \mid s_0 = s, c_{0a_0} = \mathbf{c}, \pi \right]. \quad (3)$$

Same as previous works [21, 25] on policy gradients, we denote the discounted state density by $\rho^\pi(s) = \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t P_0(s_0) P(s_0 \rightarrow s, t, \pi) ds_0$, where $P_0(s_0)$ is the probability density of initial states, and $P(s \rightarrow s', t, \pi)$ is the probability density at state s' after transitioning for t time steps from state s .

Thus we can rewrite the objective as

$$J(\pi) = \mathbb{E}_{s \sim \rho^\pi, \mathbf{c} \sim g^s, a \sim \pi(s, \mathbf{c})} [R(s, c_a)]. \quad (4)$$

3 POLICY GRADIENTS FOR ONE-STEP CONTEXTUAL RECOMMENDATIONS

In this section we investigate several key features of our proposed PGCR method. For readability, we first discuss the one-step contextual recommendation case (corresponding to section 2.1), which can be modeled as the standard contextual-bandits. Later in the next section, we will show how to extend to the generalized multi-step recommendation which can be modeled as MDP-CR.

3.1 Marginal Probability for Choosing an Item

Due to the assumption of the problem setting that the reward only depends on the selected context, we claim that for any policy π , there exists a *permutation invariant policy* that obtains at least its performance.

DEFINITION 1 (PERMUTATION INVARIANT POLICY). *A policy $\pi(\mathbf{c})$ is said to be permutation invariant if for all $\mathbf{c} \in \mathcal{C}^m$ and any its permutation $\mathbf{c}' := P(\mathbf{c})$, it has*

$$c'_{a'} \stackrel{\text{dist}}{=} c_a \quad (5)$$

where $a' \sim \pi(\mathbf{c}')$ and $a \sim \pi(\mathbf{c})$ denote for the actions chosen by the policy π , $c'_{a'}$ and c_a is the their corresponding contexts respectively, and $\stackrel{\text{dist}}{=}$ means the probability distribution of the expressions on two sides are the same.

LEMMA 1. *For any policy π , there exists a permutation invariant policy π' s.t. $J(\pi') \geq J(\pi)$.*

PROOF. Let us suppose, for the sake of contradiction, that there exists a policy π such that

(i) it is not permutation invariant, i.e. there exists $\mathbf{c} \in \mathcal{C}^m$ and some permutation operator $P \in \mathcal{P}$ that $c_a \neq c'_{a'}$ where $\mathbf{c}' = P(\mathbf{c})$ and $a' \sim \pi(\mathbf{c}')$;

(ii) The expected reward following π is larger than all permutation invariant policies $\tilde{\pi}$ that $J(\pi) > J(\tilde{\pi})$.

Then it follows that $\mathbb{E}_{\mathbf{c}} [R(\mathbf{c}_{\pi(\mathbf{c})})] > \mathbb{E}_{\mathbf{c}} [R(\mathbf{c}_{\tilde{\pi}(\mathbf{c})})]$ for all permutation invariant $\tilde{\pi}$, where the expectation is over all sets of contexts. Recall that the contexts are drawn i.i.d. from the same distribution, so we have

$$\mathbb{E}_{\mathbf{c}} [R(\mathbf{c}_{\pi(\mathbf{c})})] \equiv \mathbb{E}_{\mathbf{c}} \left[\frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} R(P(\mathbf{c})_{\pi(P(\mathbf{c}))}) \right] > \mathbb{E}_{\mathbf{c}} [R(\mathbf{c}_{\tilde{\pi}(\mathbf{c})})], \quad (6)$$

so there exists at least one \mathbf{c} that

$$\frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} R(P(\mathbf{c})_{\pi(P(\mathbf{c}))}) > R(\mathbf{c}_{\tilde{\pi}(\mathbf{c})}) \text{ for all } \tilde{\pi}. \quad (7)$$

But because π is not permutation invariant, we find a policy $\pi^*(P(\mathbf{c})) = \pi((P^* P^T P)(\mathbf{c}))$ that is permutation invariant, where

$$P^* = \arg \max_{P \in \mathcal{P}} R(P(\mathbf{c})_{\pi(P(\mathbf{c}))}),$$

then

$$R(\mathbf{c}_{\pi^*(c)}) = R(P^*(\mathbf{c})_{\pi(P^*(c))}) > \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} R(P(\mathbf{c})_{\pi(P(\mathbf{c}))}),$$

which leads to a contradictory to (6) and (7). So it must be that Lemma 1 holds. \square

Lemma 1 states that, without loss of generality, we can focus on permutation invariant policies. The objective in then becomes

$$\begin{aligned} J(\pi) &= \mathbb{E}_{\mathbf{c}} [R(c_a) \mid a \sim \pi(\mathbf{c})] \\ &= \mathbb{E}_{\mathbf{c}} \left[\sum_{i=1}^m R(c_i) I_{(a=i)} \mid a \sim \pi(\mathbf{c}) \right] \\ &= \sum_{i=1}^m \mathbb{E}_{\mathbf{c}} \left[R(c_i) I_{(a=i)} \mid a \sim \pi(\mathbf{c}) \right] \\ &= \sum_{i=1}^m \mathbb{E}_{c_i} \left[R(c_i) \mathbb{E}_{c_{-i}} [I_{(a=i)} \mid a \sim \pi(c, c_{-i})] \right] \\ &= m \mathbb{E}_{\mathbf{c}} [R(c)p(c)], \end{aligned} \quad (8)$$

where $p(c)$ is the marginal probability of choosing an item with context c (in expectation of randomness of the other $m-1$ items, denoted as c_{-1}), by a permutation invariant policy:

$$p(c) = \mathbb{E}_{c_{-1}} [I_{(a=1)} \mid a \sim \pi(c, c_{-1})]. \quad (9)$$

Suppose we have a score function μ_θ which takes the context as inputs and outputs a score, where θ are the parameters. We can construct a class \mathcal{M} of permutation invariant policies with the score function:

$$\pi_\theta(\mathbf{c}) \stackrel{\text{dist}}{=} g(\mu_\theta(c_1), \dots, \mu_\theta(c_m)), \quad (10)$$

where g is an operator that satisfies permutation invariance, for example, a family of probability distributions, and $\stackrel{\text{dist}}{=}$ means the two sides are equivalent in the sense of probability distribution.

Note that this class of policies include policies of most well-known value-based bandit algorithms. For example, if the score function is the estimation of the reward, and the operator $g(\cdot)$ chooses the item with the maximum estimated reward with probability $1-\epsilon$ and chooses randomly with probability ϵ , the policy is exactly the well-known ϵ -greedy policy [24]. If the score function is a summation of the reward estimation and the upper confidence bound, and $g(\cdot)$ chooses the item with the maximum score, it results in a similar policy to the upper confidence bound (UCB) policy [5, 17].

The policy gradient $\nabla_\theta J(\pi_\theta)$ for the standard one-step recommendations can be directly derived from (8)

$$\nabla_\theta J(\pi_\theta) = m \mathbb{E}_{\mathbf{c}} [R(c) \nabla_\theta p_\theta(c)]. \quad (11)$$

So it involves computing the marginal probabilities of choosing an item $p(c)$ which is not explicitly parameterized or tractable given an arbitrary policy π_θ . To this end, we put forward a restricted family of stochastic policies that allows us to have $p(c)$ in closed-form so as to estimate the gradient of $J(\pi_\theta)$ efficiently.

3.2 A Simple but Powerful Class of Policies

Now we propose a class of policies for our PGCR algorithm. For the standard one-step contextual-bandits, following the form of a policy described in (10), we define a class of stochastic policies denoted by \mathcal{N} as

$$\pi_\theta(\mathbf{c}) = \text{Multinoulli}\{\sigma(\mu_\theta(c_1), \dots, \mu_\theta(c_m))\}, \quad (12)$$

where $\mu_\theta(c)$ is any positive continuous function of c and θ , σ is a normalization $\sigma(\mathbf{x}) := (\frac{x_1}{\sum_i x_i}, \dots, \frac{x_m}{\sum_i x_i})$ and $\text{Multinoulli}(\cdot)$ returns a multinoulli random variable (the multinoulli distribution is also known as categorical distribution).

The form of our policy (12) generalizes several important policies in reinforcement learning. For example, when $\mu_\theta(c_i)$ is an exponential function $e^{-\beta f_\theta(c_i)}$, it reduces to the well-known *softmax policy*. If β approaches to infinity, it converges to the greedy policy that chooses the item with highest score.

For any policy $\pi_\theta \in \mathcal{N}$, the marginal probability of choosing an item can be easily derived as

$$p_\theta(c_i) = \mathbb{E}_{c_{-i}} \left[\frac{\mu_\theta(c_i)}{\mu_\theta(c_i) + \sum_{j \neq i} \mu_\theta(c_j)} \right], \quad (13)$$

which is a continuous and positive function of parameters θ . So it is straightforward to estimate $p_\theta(c)$ by sampling c_{-i} from the contexts that the player have seen before. We denote by D_t the contexts that have appeared up to step t . The estimation of $p(c)$ is unbiased because as assumed all contexts in D_t are i.i.d from the context space.

4 POLICY GRADIENTS FOR SEQUENTIAL CONTEXTUAL RECOMMENDATIONS

In this section we extend this approach to MDP-CR. We first use \tilde{c} to denote the *augmented context* by pairing together a state s and the context c for a certain item, i.e., $\tilde{c} = (s, c)$. For example, in a personalized recommender system, the state s consists of the user feature, c contains the item feature. Their concatenation $\tilde{c} = (s, c)$ is often used in ordinary tasks such as the prediction of click-through rate.

Given a policy π , the states can be roughly thought of as drawn from the discounted stationary distribution $\rho^\pi(s)$. As we already defined the density of contexts given the state as $g^s(c)$, we have the discounted density of the augmented context \tilde{c} by the axiom of probability

$$\xi^\pi(\tilde{c}) = \rho^\pi(s) g^s(c).$$

Since we assume the state distribution $\rho^\pi(s)$ is stationary, it is natural that $\xi^\pi(\tilde{c})$ is also stationary.

Then by applying the same technique as we derive the marginal probability, we derive the performance objective as follows:

$$J(\pi) = m \mathbb{E}_{\tilde{c} \sim \xi^\pi} [R(\tilde{c}) \cdot p(\tilde{c})], \quad (14)$$

where $p(\tilde{c}) = \mathbb{E}_{c_{-1} \sim g^s} [I(a=1) \mid \pi]$ is the marginal probability of choosing c given s .

Now we derive the gradients of $J(\pi)$ similar to the result in [25]. Surprisingly, it only replaces $R(c)$ in (11) by $Q(\tilde{c})$.

THEOREM 2. *Assuming the policy π leads to stationary distributions for states and contexts, the unbiased policy gradient is*

$$\nabla_\theta J(\pi_\theta) = m \mathbb{E}_{\tilde{c} \sim \xi} [\nabla_\theta p_\theta(\tilde{c}) Q(\tilde{c})], \quad (15)$$

where $Q^\pi(\tilde{c}) := Q^\pi(s, c)$ as defined in (3), and $\xi^\pi(\tilde{c})$ is the discounted density of \tilde{c} .

PROOF. We denote the state-value for state s under policy π as

$$V^\pi(s) = m \int_c p_\theta(s, c) Q^\pi(s, c) g^s(c) dc,$$

it follows that

$$\begin{aligned} \nabla_\theta V^\pi(s) &= \nabla_\theta m \int_c p_\theta(s, c) Q^\pi(s, c) g^s(c) dc \\ &= m \int_c \nabla_\theta p_\theta(s, c) Q^\pi(s, c) g^s(c) dc \\ &\quad + \gamma m \int_{s'} P(s \rightarrow s', 1, \pi) \nabla_\theta V^\pi(s') ds', \end{aligned}$$

By repeatedly unrolling the equation, we have

$$\nabla_\theta V^\pi(s) = m \int_{s'} \sum_{t=0}^{\infty} \gamma^t P(s \rightarrow s', t, \pi) \int_c \nabla_\theta p_\theta(\tilde{c}) Q(\tilde{c}) g^{s'}(c) dc ds'.$$

Integrating both side over the start-state and recalling the discounted state density $\rho^\pi(s)$ and discounted augmented context density $\xi^\pi(\tilde{c})$, we get the policy gradient as

$$\begin{aligned} \nabla_\theta J(\pi) &= m \int_s P_0(s) \nabla_\theta V^\pi(s) ds \\ &= m \int_s \rho^\pi(s) \int_c \nabla_\theta p_\theta(\tilde{c}) Q(\tilde{c}) g^s(c) dc ds \\ &= m \int_{\tilde{c}} \nabla_\theta p_\theta(\tilde{c}) Q(\tilde{c}) \xi^\pi(\tilde{c}) d\tilde{c}. \end{aligned}$$

□

Again we can find the restricted class of policies to perform efficient estimation of the gradient. For MDP-CR, it is straightforward to extend (12) to introduce states by replacing c as the augmented context \tilde{c} .

5 ACTOR-CRITIC WITH FUNCTION APPROXIMATIONS

In this section, we deliver the practical algorithm to estimate the policy gradient using function approximations. In the sample-based approximations, we write collected the reward feedback as $r(c)$ for the one-step recommendation case, and $r(\tilde{c})$ for the sequential recommendation case, as the realizations of the reward random variable $R(\cdot)$.

In conventional methods based on contextual-bandits, the most direct way to estimate the reward function is to directly apply supervised learning methods to find an estimator f_ϕ with parameter ϕ minimizing the mean squared error, i.e.,

$$\min_{\phi} \frac{1}{|D_t^{(1)}|} \sum_{c \in D_t^{(1)}} (r(c) - f_\phi(c))^2, \quad (16)$$

where $D_t^{(1)} \subset D_t$ is the set of chosen contexts and $r(c)$ is the received reward for choosing context c . It is common for most value-based contextual bandit algorithms, such as ϵ -greedy, Lin-UCB and Thompson Sampling.

However, we argue that in a policy-based perspective, this kind of off-policy supervised learning brings bias. Since our goal is to maximize the expected reward $J(\pi_\theta)$ rather than minimizing the

empirical loss as in supervised learning, the marginal probabilities of choosing an item must be taken into consideration, and the form of $f_\phi(c)$ cannot be chosen arbitrarily.

Similarly, when states and state transitions are involved in MDP-CR, we also need to find an appropriate $f_\phi(\tilde{c})$ to approximate $Q(\tilde{c})$. Since we have noted that the standard one-step contextual-bandits can be seen as a special case of the generalized MDP-CR, from now we will use the notations of MDP-CR by default to deliver the main results and algorithms.

We take similar spirit to [21, 25] and define the following *compatible* conditions, to assure that the policy gradient is orthogonal to the error in value approximation.

THEOREM 3. *The policy gradient using function approximation*

$$\nabla_\theta J(\pi_\theta) = m \int_{\tilde{c}} \nabla_\theta p_\theta(\tilde{c}) \cdot f_\phi(\tilde{c}) \xi^\pi(\tilde{c}) d\tilde{c} \quad (17)$$

is unbiased to (15) if the following conditions are satisfied:

(i) *the gradients for the value function and the policy are compatible,*

$$\nabla_\phi f_\phi(\tilde{c}) = \nabla_\theta \log p_\theta(\tilde{c}), \quad (18)$$

(ii) *the value function parameters ϕ reach a local minimum of the mean squared error over the stationary context distribution such that*

$$\nabla_\phi \mathbb{E}_{\tilde{c} \sim \xi^\pi} [p_\theta(\tilde{c}) (f_\phi(\tilde{c}) - Q^\pi(\tilde{c}))^2] = 0. \quad (19)$$

PROOF. By condition (ii), as we assumed the distribution of contexts ξ^π is stationary with respect to the policy π , it is easy to see when the conditions hold,

$$m \int_{\tilde{c}} \xi^\pi(\tilde{c}) p_\theta(\tilde{c}) [Q^\pi(\tilde{c}) - f_\phi(\tilde{c})] \nabla_\phi f_\phi(\tilde{c}) = 0.$$

Then by condition (i) we have

$$m \int_{\tilde{c}} \nabla_\theta p_\theta(\tilde{c}) [Q^\pi(\tilde{c}) - f_\phi(\tilde{c})] \xi^\pi(\tilde{c}) d\tilde{c} = 0,$$

which is the difference between (15) and (17). □

5.1 The PGCR Algorithm

We now formally propose the policy gradients algorithm for general contextual recommendations, coined by PGCR. Recall that our policy returns a Multinoulli (categorical) random variable that chooses a_t by

$$a_t \sim \text{Multinoulli} \{ \sigma(\mu_\theta(s_t, c_{t1}), \dots, \mu_\theta(s_t, c_{tm})) \}.$$

The key is to estimate the marginal expected probabilities for each item. When estimating it for some context, say c_{ti} , at some state s_t , we resample from all the previously observed contexts at the same state to get another $m - 1$ contexts.

$$\hat{p}_\theta(s_t, c_{ti}) = \frac{1}{N} \sum_n \frac{\mu_\theta(s_t, c_{ti})}{\mu_\theta(s_t, c_{ti}) + \sum_c \mu_\theta(s_t, c)} \quad (20)$$

where N ($N \geq 1$) is the number of resampling times, and c in the denominator are another $m - 1$ sampled contexts from the same state to s_t .

Similar to previous actor-critic algorithms [18], we can use Sarsa updates [24] to estimate the action-value function and then update

the policy parameters respectively by the following *policy gradients for contextual-bandits* algorithm,

$$\delta_t = r_t + \gamma f_{\phi_t}(s_{t+1}, c_{(t+1)a}) - f_{\phi_t}(s_t, c_{ta}) \quad (21)$$

$$\Delta_{\phi, t}^{\text{PGCR}} = \hat{p}_{\theta}(s_t, c_{ta}) \delta_t \nabla_{\phi} f_{\phi_t}(s_t, c_{ta}) \quad (22)$$

$$\phi_{t+1} = \phi_t + \alpha_{\phi} \Delta_{\phi, t}^{\text{PGCR}} \quad (23)$$

$$\Delta_{\theta, t}^{\text{PGCR}} = \sum_{i=1}^m \nabla_{\theta} \hat{p}_{\theta_t}(s_t, c_{ti}) f_{\phi_{t+1}}(s_t, c_{ti}) \quad (24)$$

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \Delta_{\theta, t}^{\text{PGCR}}. \quad (25)$$

In practice, the gradients can be updated on mini-batches by modern optimizers such as the Adam optimizer [14] which we already used for experiments. PGCR naturally fits to deep Reinforcement Learning and Online Learning, and techniques from these area may also be applied. Note that the algorithm can also be applied to the standard contextual bandit setting without states.

5.2 Two Useful Heuristics

5.2.1 Time-Dependent Greed. Greedy in the Limit with Infinite Exploration (GLIE), is the basic criteria desired for bandit algorithms. GLIE is to explore all the actions infinite times and then to converge to a greedy policy that reaches the global optimal reward if it runs for enough time. Value-based methods can satisfy GLIE if a positive but diminishing *exploration value* is given to all the actions. But for policy-based methods, it is not straightforward, because one cannot explicitly show the exploration level of a stochastic policy.

For PGCR, on the contrary, it is easy to have GLIE by *Time-Dependent Greed*, which applies a Time-Dependent Greed factor to the scoring function $\mu(c)$. A straightforward usage is to let $\mu_{\text{greedy}}(c; t) := \mu^{\alpha t}(c)$ where α is a pre-determined positive constant value and t is the current time-step. When $t \rightarrow \infty$, the policy tends to choose only the item with the largest score. Also the marginal probability $p_{\text{greedy}}(c; t)$ remains positive with the assumption that $\mu(c) \in (0, +\infty)$ for all context c , so any item gets an infinite chance to be explored if it runs for enough time. This technique can also apply to other policy-based RL methods as well.

5.2.2 Actor-Dropout. Directed exploration is also desired. UCB and TS methods are well-known to have directed exploration so that they automatically trade-offs between exploration and exploitation and get sub-linear total regrets. The basic insight of UCB and TS is to learn the model uncertainty during the online decision-making process and to explore the items with larger uncertainty (or higher potential to get a large reward). Often a Bayesian framework is used to model the uncertainty by estimating the posterior. However, the limitation of these methods is that assumptions and prior knowledge of reward functions are required, otherwise the posterior cannot be estimated correctly.

In light of these observations, we propose *Actor-Dropout* for PGCR to achieve directed exploration. The method is simple: to use dropout on the policy network while training and inferring. We do so because it has been theoretically and empirically justified that a neural network with dropout is mathematically equivalent to a Bayesian approximation [11]. So Actor-Dropout naturally learns the uncertainty of policies and does Monte Carlo sampling when

making decisions. Since Actor-Dropout needs no prior knowledge, it can apply to more general and complex cases than UCB and TS.

To use Actor-Dropout, in practice it is good enough for exploration to add dropout to just one layer of weights. For example, for a fully-connected actor-network, one can use dropout to the weights before the output layer, with a dropout ratio of 0.5 or 0.67. It can be understood as to train several actors and to randomly pick one at each step, so it trade-offs between exploration and exploitation since each actor learns something different from each other. We also found Actor-Dropout worth trying for other RL or Online Learning tasks in the exploration phase.

5.3 Lower Variance of the Gradients of PGCR than Vanilla PG

We prove that the variance of updating the actor and the critic of PGCR is less than that of vanilla PG.

Since the concept of *context* does not exist in the classic formulation of reinforcement learning, it is often regarded as part of the state. Given a stochastic policy $\pi_{\theta}(s, c)$, PG has policy gradients

$$\nabla_{\theta} J(\pi_{\theta}) = m \sum_s \rho_s^{\pi} \sum_{i=1}^m \nabla_{\theta} [e_i^T \pi_{\theta}(s, c)] \cdot f_{\phi}(\tilde{c}_i), \quad (26)$$

where e_i denotes a unit vector and $e_i^T \pi_{\theta}(s, c)$ is the probability for choosing the i^{th} item. For simplicity, we write $v_i := e_i^T \pi_{\theta}(s, c)$. Since we focus on policy gradients, we assume that PG has a critic function $f_{\phi}(\tilde{c})$ with the same form as PGCR. The corresponding update steps for PG is

$$\Delta_{\phi_t}^{\text{PG}} = v_{ta} \delta_t \nabla_{\phi} f_{\phi}(s_t, c_{ta}) \quad (27)$$

$$\Delta_{\theta_t}^{\text{PG}} = \sum_i^m \nabla_{\theta} v_{ti} f_{\phi_{t+1}}(s_t, c_{ti}). \quad (28)$$

Our PGCR can achieve lower estimation variance comparing to classic stochastic policy gradient methods such as [25], the reasons are two-fold. Firstly, by Lemma 1 we know permutation invariant policies are sufficient for contextual-bandits problems, so PGCR adopts the restricted class of policies. On the contrary, in vanilla policy gradients, one should treat a state s and the whole contexts c altogether as the input of the policy function, so the sample space can be much larger, which results in lower sample efficiency. Secondly, even if with the same form of policy, vanilla policy gradients tend to converge slower than PGCR because they do not take the marginal expected probabilities of choosing an item into consideration.

In a formal way, we can have the following conclusion.

LEMMA 4. *Given a policy $\pi_{\theta} \in \mathcal{N}$ and a value approximation f_{ϕ} , both $\Delta_{\phi_t}^{\text{PGCR}}$ and $\Delta_{\phi_t}^{\text{PG}}$ are unbiased estimators for the true gradients of action-value approximation*

$$\Delta_{\phi_t} = p(s_t, c_{ta}) \delta_t \nabla_{\phi} f_{\phi}(s_t, c_{ta}). \quad (29)$$

And $\text{Var}[\Delta_{\phi_t}^{\text{PGCR}}] \leq \text{Var}[\Delta_{\phi_t}^{\text{PG}}]$. Additionally if PGCR uses a fixed N , as $t \rightarrow +\infty$, with probability 1 we have

$$\text{Var}[\Delta_{\phi_t}^{\text{PGCR}}] \rightarrow \frac{1}{N} \text{Var}[\Delta_{\phi_t}^{\text{PG}}]. \quad (30)$$

PROOF. It is obvious that both v_{ta} in $\Delta_{\phi_t}^{PG}$ and $\hat{p}(s_t, c_{ta})$ in $\Delta_{\phi_t}^{PGCR}$ are unbiased to $p(s_t, c_{ta})$. So both $\Delta_{\phi_t}^{PGCR}$ and $\Delta_{\phi_t}^{PG}$ are unbiased to Δ_{ϕ_t} . To analyze the variance, we focus on the estimations of the probability of choosing an item: v_{ta} and $\hat{p}(s_t, c_{ta})$. Let $V := \text{Var}[v_{ta}]$. Then for PGCR,

$$\text{Var}[\hat{p}(s_t, c_{ta})] = \text{Var}\left[\frac{1}{N} \sum_{n=1}^N v_{ta}^{(n)}\right],$$

where $v_{ta}^{(n)}$ denotes the probability of choosing c_{ta} at the n^{th} time of sampling. In the worst case, it samples exactly the same set of $m-1$ items every time, then $\text{Var}[\hat{p}(s_t, c_{ta})] = V$. Otherwise if there exists n_1 and n_2 that the samples are different such that $v_{ta}^{(n_1)} \neq v_{ta}^{(n_2)}$, then the correlation is strictly less than 1 and we have $\text{Var}[\Delta_{\phi_t}^{PGCR}] < \text{Var}[\Delta_{\phi_t}^{PG}]$ in this case. Finally when enough time steps passed, for N is a fixed positive integer, the probability of each item being sampled at most once is

$$\binom{mt}{(m-1)N} (mt)^{-(m-1)N} \rightarrow 1 \quad \text{as } t \rightarrow +\infty.$$

So with probability 1 the sampled contexts are all different to each other so the estimated probabilities of choosing an item are i.i.d., then $\text{Var}[\Delta_{\phi_t}^{PGCR}] \rightarrow V/N$. \square

We get the following theorem applying the similar technique to Lemma 4. We claim that Policy gradients (17) has no higher variance than gradients in PG.

THEOREM 5. $\text{Var}[\Delta_{\theta}^{PGCR}] \leq \text{Var}[\Delta_{\theta}^{PG}]$.

PROOF. Similar to the proof of Lemma 4, we denote the variance of Δ_{θ}^{PG} by V_{θ} .

$$V_{\theta} := \text{Var}[\Delta_{\theta}^{PG}] = \text{Var}\left[\sum_i^m \nabla_{\theta} v_{ti} f_{\phi_{t+1}}(s_t, c_{ti})\right]$$

By the update rules (24) of PGCR, the variance of Δ_{θ}^{PGCR} is

$$\begin{aligned} \text{Var}[\Delta_{\theta}^{PGCR}] &= \text{Var}\left[\sum_{i=1}^m \nabla_{\theta} \hat{p}(s_t, c_{ti}) f_{\phi_{t+1}}(s_t, c_{ti})\right] \\ &= \text{Var}\left[\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^m \nabla_{\theta} v_{ti}^{(n)} f_{\phi_{t+1}}(s_t, c_{ti})\right] \\ &\leq \text{Var}\left[\sum_{i=1}^m \nabla_{\theta} v_{ti}^{(n)} f_{\phi_{t+1}}(s_t, c_{ti})\right], \quad \forall n = 1, \dots, N. \end{aligned}$$

Because of the assumption that the sampled contexts in each sampling procedure are independent and identical distributed, we have $\text{Var}[\sum_{i=1}^m \nabla_{\theta} v_{ti}^{(n)} f_{\phi_{t+1}}(s_t, c_{ti})] = V_{\theta}$ for all $n = 1, \dots, N$ and the theorem is proved. \square

Note that, in practice, PGCR does not necessarily set N to a large integer since it is naturally a finer-grained experience replay [3]. Surprisingly, when $N = 1$, PGCR can have a better performance than PG even in the simplest setting. In the next section, we will demonstrate experimental results that show that PGCR with $N = 1$ achieves better performance in various settings compared to other baseline methods including PG.

The results can be interpreted as follows. From a statistical point of view, PGCR takes advantage from a resampling technique so the estimations have lower variances. From an optimization perspective, PGCR reduces the correlation of estimating probabilities of choosing the m items within the same time step, so it has less chance to suffer from exploiting and over-fitting, while vanilla PG cannot. For example, when the estimated values of m contexts are given, an optimizer for PG would simultaneously increase one item's chosen probability and reduce other $m-1$ ones', which results in training the policy into a deterministic one: the item with the largest estimated value will get a chosen probability close to 1, and others get arbitrary small probabilities close to 0. Afterward, the items with 0 chosen probabilities will hardly have any influence to further updates. So eventually, vanilla PG is likely to over-fit the existing data. On the contrary, when PGCR estimates the gradients, even if an item cannot beat against the other $m-1$ competitors at its own time step, it can still help because it outranked some items from other time steps. Therefore, PGCR tends to be more robust and explores better than vanilla PG.

6 EXPERIMENTS

6.1 Datasets and Simulation Details

We test our proposed PGCR and other baseline methods on several simulated contextual recommendation environments. To start with, we would like to describe the details of our simulations.

- **Toy environments:** The first set of experiments are done with a generated toy dataset to simulate the standard contextual bandit settings. We simulate a contextual bandit environment with 5 items at each step, where each item is represented by a 40-dimensional context vector that is i.i.d. sampled from a uniform distribution in a unit cube $c \sim U(C)$, $C = (0, 1)^{40}$. Once an item with context c is chosen by the player, the environment returns a random reward $R(c)$. We test three types of reward functions: (a) the linear reward with Gaussian noise, as $R(c) := w_r^T c + e_r$; (b) the Bernoulli reward, as $R(c) \sim \text{Bernoulli}(\beta(c))$ where $\beta(c) := w_{\beta}^T c + e_{\beta} \in [0, 1]$ is the probability to return reward 1 for choosing c ; (c) the mixed reward, which first returns a random linear reward $w_r^T c + e_r$ with probability $\beta(c)$ and returns a zero reward with probability $1 - \beta(c)$, as a mixture of binary and linear rewards. w_r and w_{β} are coefficients unknown to the player. e_r and e_{β} are white noises to introduce some randomness.
- **Music recommendation environments:** We use a real-world dataset of music recommendation provided by *KKBox* and open-sourced on *Kaggle.com*¹. The challenge is to predict the chances of a user listening to a song repeatedly. So we construct simulation environments to simulate the online contextual recommendations to test our methods. We construct two simulators based on the distributions of the dataset with different settings: one simplified setting without explicit states and one general setting with states and state transitions. At each time step, a user comes to the system, who is randomly picked from the users in the dataset.

¹<https://www.kaggle.com/c/kkbox-music-recommendation-challenge/>

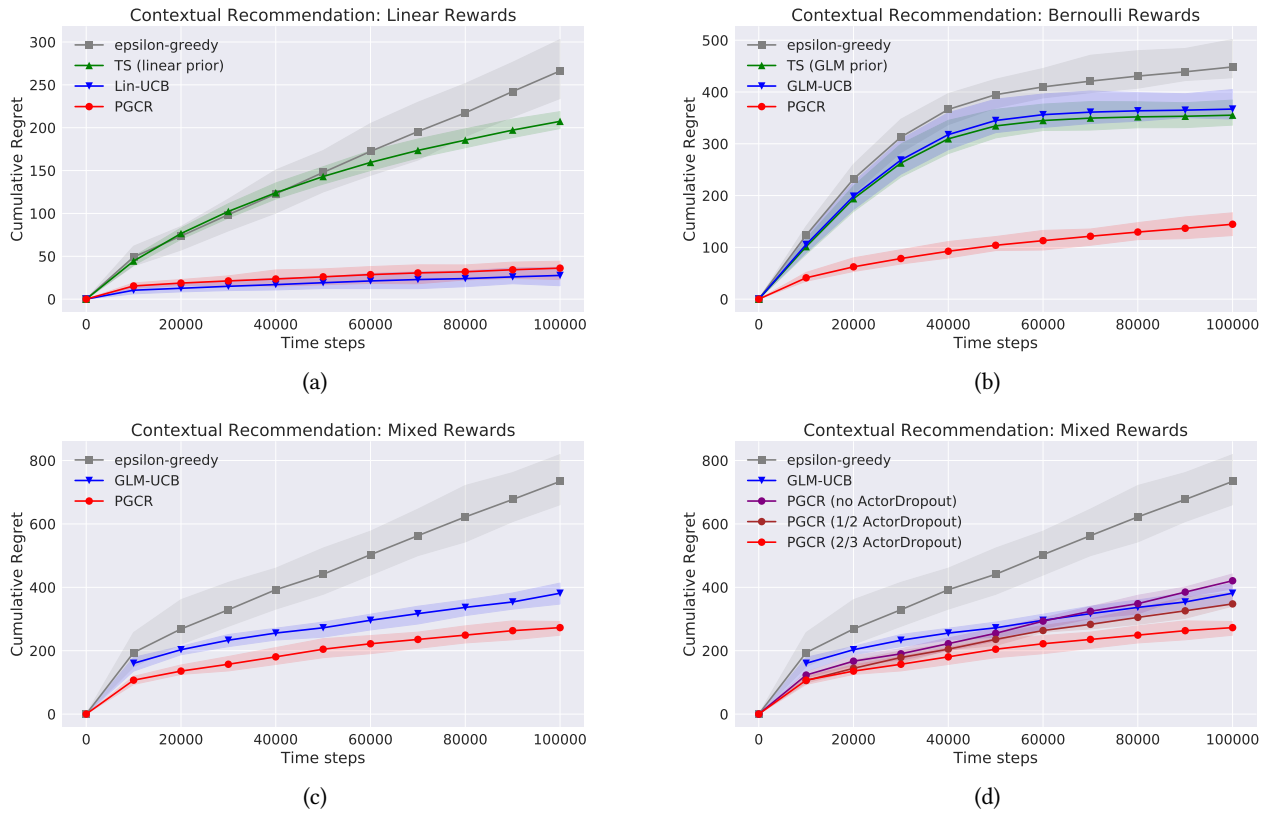


Figure 1: Experiments on toy data. The solid lines are averaged cumulative regrets over 20 runs, and the shaded areas stand for the standard deviations. (a) Linear rewards: PGCR perform comparably to Lin-UCB and is much better than the other two. (b) Bernoulli rewards: PGCR outperforms the others with large margins. (c) Mixed-rewards: PGCR outperforms the others. (d) Testing PGCR with different level of Actor-Dropout. It shows that, without Actor-Dropout, PGCR tends to have linear regret like ϵ -greedy and fail to converge. But when using Actor-Dropout, PGCR empirically converges which is similar to UCB, yet gets even lower regret than UCB.

For the state-aware setting, we set the last 3 songs the system recommended previously to this user together with the corresponding feedbacks (listened or not) as his/her current state. Then the simulator randomly samples a set of 10 candidate songs from the user’s listening history. Finally, the recommender needs to make a decision to recommends one to the user. If the user listens to it again (which is the original target for the supervised learning dataset), the system gets a reward 1, otherwise, it gets a reward 0. Each song has 94 fields of features in the context vector, including discrete attributes and numerical features about the song’s genre, artists, composers, language, etc. The simulation consists of 5 million time steps and is repeated for 5 runs. Since the optimal policy is unknown in this problem, we will use the average reward as the performance metric.

6.2 Experiments on Toy Datasets

For the three standard contextual-bandit problems, we use the cumulative regret as the evaluation metric. The cumulative regret

is defined as the cumulative difference between the received reward and the reward of the optimal item.

For PGCR, we use the multi-layer perceptron (MLP) as the policy and value networks. The MLPs have one fully-connected hidden layer with 10 units and the ReLU activation. As for the training details, at each step, we sample a batch of 64 history samples for efficient experience replay. The loss is minimized by the gradient-based Adam optimizer [14].

We compare PGCR with the following algorithms:

- **ϵ -greedy:** It chooses the item with the largest estimated value with a probability of $1 - \epsilon$ and chooses randomly otherwise. Specifically, it estimates the expected reward by a value network with the same structure of PGCR.
- **Lin-UCB:** The widely studied version of UCB for contextual bandits with linear rewards [1, 9, 17], which uses a linear function to approximate the reward, and chooses the item with the maximum sum of the estimated reward and the estimated confidence bound.
- **GLM-UCB:** The UCB method for generalized linear rewards, proposed in [10], which can solve non-linear rewards if the

reward function can be fitted by a generalized linear model of contexts, such as Bernoulli rewards and Poisson rewards.

- **Thompson Sampling:** It samples from the posterior distribution of parameters, estimates each item’s value, and chooses the item with the maximum estimation [4, 8]. Specifically, it uses the same function approximation as Lin-UCB or GLM-UCB for linear and non-linear rewards.

For Lin-UCB, GLM-UCB and Thompson Sampling, we use the same training procedures as suggested in [17]. For ϵ -greedy, ϵ is set constantly to 0.1.

For each method, in order to reduce the randomness of experiments, we run the simulation for 20 times and report their averaged cumulative regrets. The results are shown in Figure 1.

From Figure 1(a), (b) and (c), it sees that PGCR converges more quickly and has lower regrets in most cases. For the linear bandits, Lin-UCB is, theoretically, one of the best choice if the linearity is known explicitly beforehand. Even though PGCR does not know the form of reward function apriori, the performance is comparable with Lin-UCB. Thompson Sampling converges slower and ϵ -greedy obviously fails to converge. For the generalized linear case, PGCR achieves much lower regret than other baselines. For the mixed reward case, we see that PGCR learns faster than GLM-UCB and empirically converges after a long run.

Furthermore, we did an ablation study to test the improvements induced by the proposed heuristic, Actor-Dropout. See Figure 1(d) for the comparisons.

It shows that Actor-Dropout significantly helps PGCR to converge. The growth rate of cumulative regret for PGCR without Actor-Dropout is similar to ϵ -greedy, indicating that the original algorithm fails to converge and has linear regrets. But when equipped with Actor-Dropout, the regrets are smaller. When the dropout rate is set to 0.67, the growth rate of PGCR’s regret is similar to that of GLM-UCB which can theoretically achieve sub-linear regrets. So empirically we remark that Actor-Dropout is a strong weapon for PGCR in order to get a convergence guarantee, even with almost no assumptions on the problem.

6.3 Experiments on Music Recommendations

6.3.1 Simplified setting without explicit states. The experimental setup in the setting without states is as follows: PGCR uses MLPs as policy and value networks. Each MLP has two hidden layers of sizes 60 and 20 respectively and uses the ReLU nonlinearity. ϵ -greedy has exactly the same value network like the one in PGCR. All these methods are trained with Adam algorithm with the same learning rates. The batch-size is set as 256. We also test GLM-UCB here as another baseline because the reward is binary thus can be learned by a logistic regression model.

As is shown in Figure 2(a), PGCR performs the best. It is interesting to see that traditional contextual-bandits methods learn fast from the beginning, which indicates that they are good at exploration, but their average rewards stop increasing rapidly due to the limitation of the fitting power of general linear models. ϵ -greedy learns slowly at the early stage, but it outperforms GLM-UCB and TS after a long run. Comparing with these algorithms, PGCR has the best performance from the beginning to the end of the learning process.

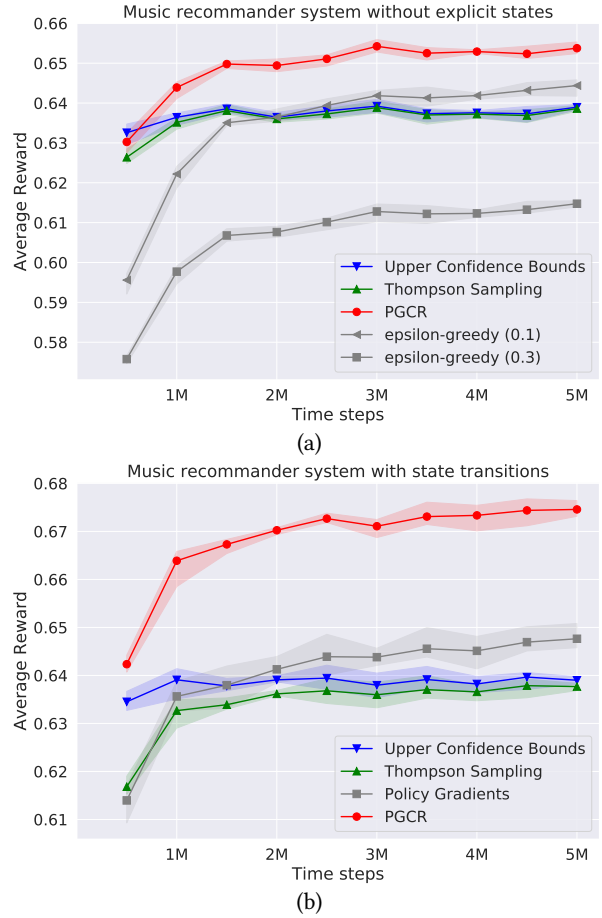


Figure 2: Average rewards of episodes for (a) music recommender without explicit states; (b) music recommender with states and state transitions. The solid lines are averaged from 5 runs, and the shaded areas are standard deviations.

6.3.2 Simulation with states and transitions. The experimental setup for this experiment with states is as follows. We enlarge the size of the first hidden layer of the MLPs from 60 to 90 because there are more inputs (including the contexts and the states). UCB and TS here take the augmented contexts as input. They keep the same general linear modeled priors as the previous part.

The result of the experiment is shown in figure 2(b). PGCR outperforms other algorithms with larger map comparing with the previous experiment. An interesting fact is that both UCB and TS can only get almost the same return as in the previous experiment, which indicates that they can hardly make any use of the information from the states. PGCR learns faster and gets state-of-the-art performance in this task.

Since there are states and transitions, we also test the vanilla Policy Gradient method [25] as another baseline, with the same MLP neural networks as actor and critic as PGCR. From the result, it shows that vanilla PG can outperform classic bandits methods, which is not surprising because it can make use of the state dynamics and maximize the long-term return. However, it is still much

worse than our purposed PGCR, for the reason that PGCR has a smaller search space and smaller variance when estimating the policy gradients. So the experimental results verified that PGCR is more sample efficient than the vanilla policy gradient method.

These simulation experiment results indicate that PGCR provides an alternative to conventional methods by using stochastic policies which can address the trade-off between exploration and exploitation well. The fast learning in the beginning phase and the stable performance over the entire training period of PGCR show that it is reliable to apply for real-world recommender systems.

7 CONCLUSION AND DISCUSSION

This paper has studied how to use the actor-critic algorithm with neural networks for general contextual recommendations, without unrealistic assumptions or prior knowledge to the problem. We first show that the class of permutation invariant policies is sufficient for our problem, and then derive the expected return of a policy depends on its marginal expected probability of choosing each item. We next propose a restricted class of policies in which the objective has a simple closed form and is differentiable to parameters. We prove that when using policies in this class, the gradient can be computed in closed-form. Furthermore, we propose Time-Dependent Greed and Actor-Dropout to significantly improve the performance and to guarantee the convergence property. Eventually, it comes to our proposed PGCR algorithm. The algorithm can be applied to standard contextual bandits as well as the generalized sequential decision-making problems with state and state transitions.

By testing on a toy dataset and a recommendation dataset, we showed that PGCR indeed achieves state-of-the-art performance for both classic one-step recommendations and MDP-CR with state transitions in a real-world scenario.

It is a promising direction for the future work to extend our results to a variant of realistic recommendation settings, i.e, online advertising systems that choose multiple items at each step, or learning-to-rank that carries out diverse recommendation.

ACKNOWLEDGEMENT

The work was supported by the National Key Research and Development Program of China under Grant No. 2018YFB1004300, the National Natural Science Foundation of China under Grant No. U1836206, U1811461, 61773361, the Project of Youth Innovation Promotion Association CAS under Grant No. 2017146.

Pingzhong Tang and Qingpeng Cai were supported in part by the National Natural Science Foundation of China Grant 61561146398, a China Youth 1000-talent program and an Alibaba Innovative Research program.

REFERENCES

- [1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*. 2312–2320.
- [2] Naoki Abe, Alan W Biermann, and Philip M Long. 2003. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica* 37, 4 (2003), 263–293.
- [3] Sander Adam, Lucian Busoni, and Robert Babuska. 2012. Experience replay for real-time reinforcement learning control. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 2 (2012), 201–212.
- [4] Shipra Agrawal and Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*. 127–135.
- [5] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.
- [6] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. 2012. A contextual-bandit algorithm for mobile context-aware recommender system. In *International Conference on Neural Information Processing*. Springer, 324–331.
- [7] Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. 2018. Reinforcement Mechanism Design for e-commerce. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1339–1348.
- [8] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.
- [9] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 208–214.
- [10] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*. 586–594.
- [11] Yariv Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. 1050–1059.
- [12] Nicolas Heess, Jonathan J Hunt, Timothy P Lillicrap, and David Silver. 2015. Memory-based control with recurrent neural networks. *arXiv preprint arXiv:1512.04455* (2015).
- [13] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11, Apr (2010), 1563–1600.
- [14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [15] Andreas Krause and Cheng S Ong. 2011. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*. 2447–2455.
- [16] John Langford and Tong Zhang. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*. 817–824.
- [17] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.
- [18] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [19] Benedict C May, Nathan Korda, Anthony Lee, and David S Leslie. 2012. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* 13, Jun (2012), 2069–2106.
- [20] Guy Shani, David Heckerman, and Ronen I Brafman. 2005. An MDP-based recommender system. *Journal of Machine Learning Research* 6, Sep (2005), 1265–1295.
- [21] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *ICML*.
- [22] Aleksandr Slivkins, Filip Radlinski, and Sreenivas Gollapudi. 2013. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *Journal of Machine Learning Research* 14, Feb (2013), 399–436.
- [23] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. 2012. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory* 58, 5 (2012), 3250–3265.
- [24] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.
- [25] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*. 1057–1063.
- [26] Nima Taghipour and Ahmad Kardan. 2008. A hybrid web recommender system based on q-learning. In *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 1164–1168.
- [27] Liang Tang, Yexi Jiang, Lei Li, and Tao Li. 2014. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 73–80.
- [28] Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. 2015. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 323–332.
- [29] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. 2013. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1587–1594.
- [30] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.
- [31] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* 25, A (1988), 287–298.