On reviewer 1’s concern about the general example and general reward functions

A contextual example that can not be covered if we just extend existing results to involve contextual information is given in Section 4.3 and is experimented in Section 5.3 with Figure 3(c).

This example considers a network routing problem with latency. The latency of each edge in the network follows an exponential distribution, whose mean is related to the features of the corresponding edge. It is simply assumed that such relations are all linear combinations. We regard an edge is blocked if its latency is larger than some tolerance \tau. Usually we will not wait until the edges to respond and we will measure the edges’ latencies until some time, say 2\tau. That is, what we measure for an edge is indeed the samples of a cut-off exponential distribution whose original mean is a linear combination of its features. And the probability of an edge to be blocked is probability of a cut-off exponential random sample is larger than the tolerance \tau and the probability of an edge to be unblocked is the probability of the cut-off exponential random sample is less than the tolerance \tau. Therefore, a path is unblocked if and only if all its edges are unblocked, and the probability is the product of each edge of the path being unblocked. If we only extend the existing works to involve contextual information, this example can not be covered. This example is also checked in Section 5.3 with Figure 3(c)

On reviewers’ concern about experimental results

When we demonstrate the experimental results, we focus on the comparison aspect to show the advantage of involving contextual information and position discounts. Actually the learning curves will show a good concave/convex shape when $n$, number of rounds, is small. Taking Figure 1(b) for an example, the learning curves show a good concave shape when $1 \leq n \leq 2000$, during which the estimation for theta has a cosine similarity of 0.995, that is $cos(\theta, \hat{\theta}) = 0.995$. Then due to the randomness, the accuracy of $\hat{\theta}$ increases slowly. So there is a very small probability that $\hat{\theta}$ will choose the wrong action, or super arm. Such a small probability is decreasing, but in a very slow rate, which reflects the shapes of the curves. The learning curves will be better if we demonstrate the prominent part ($n$ small, but this case is not convincing) or we use big $n$(but in this situation, the comparisons will look trivial because our regrets will be relatively very small like Figure 1(a,c)).

On reviewer 1’s concern about monotonicity assumption

This assumption implies if the recommended list has higher reward means, then the expected reward of the list should also get higher. We think this is a relatively basic assumption.

On reviewer 1’s concern about the name C^3 UCB

We first wanted to use the name ContCombCascade-UCB, but this seems a bit longer. For the limited length, we use the abbreviative C^3-UCB instead. We can introduce in our paper its full name first and then use its abbreviation.

On reviewer 1’s concern about last paragraph in Section 5.3

The last paragraph is related with Figure 3(d) which demonstrates the benefits to adopt the position discounts when the true setting has one. When applying algorithms to real applications, usually we are not 100% sure about the compatibility. Suppose the true criterion has a position discount $\gamma^\ast = 0.9$. Then the algorithm with no position discounts (or $\gamma = 1$) doesn’t fit the setting well and learns in a wrong direction. The regret of using the true $gamma = 0.9$ is only about 13% of that using no position discounts ($\gamma = 1$).

On reviewer 2’s concern about Line 657-663

This part is mainly for the comparisons of our results with existing results. Our result, when restricted to (Kveton et al., 2015c), has an additional term. The reason is we use the main result of the linear bandits (Abbasi-Yadkori et al. 2011). If the setting of linear bandits is restricted to MAB setting, the bound can match the lower bound of (Lai & Robbins 1985). Our results build on this paper, so if our setting is restricted to (Kveton et al., 2015c), it is similar to restrict the linear bandits setting to the MAB setting. So we will indeed have same regret bounds. Details are omitted in this part. We can add the details in the Appendix to make it clear.

On reviewer 2’s concern about the context in Section 5.3

Because we don’t have a real data for this problem except the network structure, we generate the contextual information x\_{t,a} randomly. C^3-UCB has better performance than CombCascade because they don’t make use of the context information.

On reviewers’ other concerns

The V\_t norm in line 376 is defined in line 362-364. Due to the length limit, we didn’t bring the definition out. Sorry for the inconvenience.

The simple path in Line 812 is a directed path without cycles. We will make it clear.

f^\ast is only defined in Theorem 4.6 because we only need it there. The f\_^\ast is defined in Line 328, which is more essential.