We thank all the reviewers for their constructive comments. We address the main points of the reviews in the following. We will also address other comments in the revised version. We abbreviate Thompson sampling and partial monitoring to TS and PM, respectively.

To Reviewers #1, #2, and #3 (on Generalization to Sub-Gaussian Noise)

The restriction to the Gaussian noise comes from the essential difficulty of the problem-dependent analysis of TS, where lower bounds for some probabilities are needed whereas the sub-Gaussian assumption is suited for obtaining upper bounds. In fact, to the best of our knowledge, the problem-dependent regret analysis for TS on the sub-Gaussian case has never been investigated even for the multi-armed bandit setting, which is quite simple compared to that of PM. In the literature, the noise distribution is restricted to distributions with explicitly given forms, e.g., Bernoulli, Gaussian, or more generally a one-dimensional canonical exponential family (Kaufmann et al., 2012; Agrawal and Goyal, 2013a, Korda et al., 2013). Their analysis relies on the specific characteristic of the distribution to bound the problem-dependent regret. We will add this discussion in the revised version.

(Korda et al., 2013) Thompson Sampling for 1-Dimensional Exponential Family Bandits, In NeurIPS2013.

To Reviewer #2

> "Why not make the entire focus of the paper the setting where theoretical results can be obtained?"

The linear PM has been mainly considered from the theoretical viewpoint and experiments have not been conducted. Therefore, we conducted experiments on the discrete setting for fair comparison with existing work.

> "In line 216 it is unclear whether the algorithm is different from the family of TS used in practical settings"

Their algorithm considers the posterior distribution for *regret* (not pseudo-regret), and an action is chosen according to the posterior probability that each arm minimizes the *cumulative* regret (thus the time horizon also needs to be known), whereas the typical TS considers the pseudo-regret at each round. We will make it more clear in the revised version.

> "The experimental section ... incompleteness by not acknowledging this approach (= Mario sampling)."

Thank you for pointing out the lack of reference to the important work. Mario sampling is the algorithm for locally observable games and not applicable to hard games, like dp-hard. On the other hand for locally observable games, Mario sampling coincides with TS (except for the above difference between pseudo-regret and regret with known time horizon) when any pair of actions is a neighbor. We confirmed that some dp-easy games satisfy this property, and conjecture that it generally holds for dp-easy. Therefore, the performance is essentially the same between TSPM ($R = 1$) and Mario sampling, though general analysis on the difference is an important future direction.

> "In the comments on the upper bound it would be useful to have some sense of the magnitude of the $z_{j,k}$ terms."

Intuitively, the norm of $z_{j,k}$ indicates the difficulty of the problem. Whereas we can estimate $(S_j p, S_k p)$ with noise through taking actions j and k, the actual interest is the gap of the losses $p^\top(L_j - L_k) = (S_j p, S_k p)^\top z_{j,k}$. Thus, if $\|z_{j,k}\|$ is large, the gap estimation becomes difficult since the noise is enhanced through $z_{j,k}$. We will add this discussion in the revised version.

To Reviewer #3 and Reviewer #4 (on the Number of Rejected Times in Accept-Reject Sampling)

In the accept-reject sampling, it is desirable that the frequency of rejection (a) does not increase as the time-step t and (b) does not increase so much with the number of outcomes. From the experimental results, we can see that the property (a) is indeed satisfied. For the property (b), it is true that the frequency of rejection becomes large when exact sampling ($R = 1$) is conducted, as pointed by R3. Still, we can substantially improve this frequency by setting R to be a small value or zero, which still keeps regret tremendously better than that of BPM with almost the same time-efficiency as BPM-TS. This result exhibits a clear speed-performance trade-off. We will make it more clear in the revised version.

To Reviewer #4

> "... The paper includes experiments, but they are limited in scale. The lack of contextual feature modeling ... "

The scale of the experiment is determined based on standard literature (Bartók et al., 2012; Komiyama et al., 2015). The analysis of the contextual and the non-contextual settings is essentially different, because achievable regrets and appropriate algorithms can be different between them.

> "Experiments setup. Needs to remind reader what is N and M. (I guess number of arms?) How can they be different?"

Thank you for the suggestion. In the revised version, we will explicitly describe that “price” and “evaluation value” correspond to the action and the outcome, respectively.

> "There is a general lack of discussions of the results - how different conditions impact the cumulative regrets."

The discussion on the performance comparison of methods and the rejection sampling is given in Line 297–306. In addition to that, we can say that the proposed methods outperform BPM-TS more significantly for a larger number of outcomes. This can be seen from the discussion in Appendix D, and we will make it more clear in the revised version.