Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring

Taira Tsuchiya
The University of Tokyo
RIKEN AIP
tsuchiya@ms.k.u-tokyo.ac.jp

Junya Honda
The University of Tokyo
RIKEN AIP
honda@edu.k.u-tokyo.ac.jp

Masashi Sugiyama
RIKEN AIP
The University of Tokyo
sugi@k.u-tokyo.ac.jp

Abstract

We investigate finite stochastic partial monitoring, which is a general model for sequential learning with limited feedback. While Thompson sampling is one of the most promising algorithms on a variety of online decision-making problems, its properties for stochastic partial monitoring have not been theoretically investigated, and the existing algorithm relies on a heuristic approximation of the posterior distribution. To mitigate these problems, we present a novel Thompson-sampling-based algorithm, which enables us to exactly sample the target parameter from the posterior distribution. Besides, we prove that the new algorithm achieves the logarithmic *problem-dependent expected pseudo-regret* $\mathrm{O}(\log T)$ for a linearized variant of the problem with local observability. This result is the first regret bound of Thompson sampling for partial monitoring, which also becomes the first logarithmic regret bound of Thompson sampling for linear bandits.

1 Introduction

Partial monitoring (PM) is a general sequential decision-making problem with limited feedback (Rustichini, 1999; Piccolboni and Schindelhauer, 2001). PM is attracting broad interest because it includes a wide range of problems such as the multi-armed bandit problem (Lai and Robbins, 1985), a linear optimization problem with full or bandit feedback (Zinkevich, 2003; Dani et al., 2008), dynamic pricing (Kleinberg and Leighton, 2003), and label efficient prediction (Cesa-Bianchi et al., 2005).

A PM game can be seen as a sequential game that is played by two players: a learner and an opponent. At every round, the learner chooses an action, while the opponent chooses an outcome. Then, the learner suffers an unobserved loss and receives a feedback symbol, both of which are determined from the selected action and outcome. The main characteristic of this game is that the learner cannot directly observe the outcome and loss. The goal of the learner is to minimize his/her cumulative loss over all rounds. The performance of the learner is evaluated by the regret, which is defined as the difference between the cumulative losses of the learner and the optimal action (*i.e.,* the action whose expected loss is the smallest).

There are mainly two types of PM games, which are the *stochastic* and *adversarial* settings (Piccolboni and Schindelhauer, 2001; Bartók et al., 2011). In the stochastic setting, the outcome at each round is determined from the *opponent's strategy*, which is a probability vector over the opponent's possible choices. On the other hand, in the adversarial setting, the outcomes are arbitrarily decided by the

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

opponent. We refer to the PM game with finite actions and finite outcomes as a *finite* PM game. In this paper, we focus on the stochastic finite game.

One of the first algorithms for PM was considered by Piccolboni and Schindelhauer (2001). They proposed the FeedExp3 algorithm, the key idea of which is to use an unbiased estimator of the losses. They showed that the FeedExp3 algorithm attains $\tilde{O}(T^{3=4})$ minimax regret for a certain class of PM games, and showed that any algorithm suffers linear minimax regret $\Omega(T)$ for the other class. Here T is the time horizon and the notation $\tilde{O}(\ )$ hides polylogarithmic factors. The upper bound $\tilde{O}(T^{3=4})$ is later improved by Cesa-Bianchi et al. (2006) to $O(T^{2=3})$, and they also provided a game with a matching lower bound.

In the seminal paper by Bartók et al. (2011), they classified PM games into four classes based on their minimax regrets. To be more specific, they classified games into trivial, easy, hard, and hopeless games, where their minimax regrets are 0, $\tilde{\Theta}(\sqrt{T})$, $\Theta(T^{2=3})$, and $\Theta(T)$, respectively. Note that the easy game is also called a *locally observable* game. After their work, several algorithms have been proposed for the finite PM problem (Bartók et al., 2012; Vanchinathan et al., 2014; Komiyama et al., 2015). For the problem-dependent regret analysis, Komiyama et al. (2015) proposed an algorithm that achieves $O(\log T)$ regret with the optimal constant factor. However, it requires to solve a time-consuming optimization problem with infinitely many constraints at each round. In addition, this algorithm relies on the forced exploration to achieve the optimality, which makes the empirical performance near-optimal only after prohibitively many rounds, say, 10^5 or 10^6.

Thompson sampling (TS, Thompson, 1933) is one of the most promising algorithms on a variety of online decision-making problems such as the multi-armed bandit (Lai and Robbins, 1985) and the linear bandit (Agrawal and Goyal, 2013b), and the effectiveness of TS has been investigated both empirically (Chapelle and Li, 2011) and theoretically (Kaufmann et al., 2012; Agrawal and Goyal, 2013a; Honda and Takemura, 2014). In the literature on PM, Vanchinathan et al. (2014) proposed a TS-based algorithm called BPM-TS (Bayes-update for PM based on TS) for stochastic PM, which empirically achieved state-of-the-art performance. Their algorithm uses Gaussian approximation to handle the complicated posterior distribution of the opponent's strategy. However, this approximation is somewhat heuristic and can degrade the empirical performance due to the discrepancy from the exact posterior distribution. Furthermore, no theoretical guarantee is provided for BPM-TS.

Our goals are to establish a new TS-based algorithm for stochastic PM, which allows us to sample the opponent's strategy parameter from the exact posterior distribution, and investigate whether the TS-based algorithm can achive sub-linear regret in stochastic PM. Using the accept-reject sampling, we propose a new TS-based algorithm for PM (TSPM), which is equipped with a numerical scheme to obtain a posterior sample from the complicated posterior distribution. We derive a logarithmic regret upper bound $O(\log T)$ for the proposed algorithm on the locally observable game under a linearized variant of the problem. This is the first regret bound for TS on the locally observable game. Moreover, our setting includes the linear bandit problem and our result is also the first logarithmic expected regret bound of TS for the linear bandit, whereas a high-probability bound was provided, for example, in Agrawal and Goyal (2013b). Finally, we compare the performance of TSPM with existing algorithms in numerical experiments, and show that TSPM outperforms existing algorithms.

2 Preliminaries

This paper studies finite stochastic PM games (Bartók et al., 2011). A PM game with N actions and M outcomes is defined by a pair of a loss matrix $\mathsf{L} = (\ell_{i;j})\ 2\ \mathsf{R}^{N\ \ M}$ and feedback matrix $\mathsf{H} = (h_{i;j})\ 2\ [A]^{N\ \ M}$, where A is the number of feedback symbols and $[A] = \mathsf{f}1, 2, \ldots, A\mathsf{g}$.

A PM game can be seen as a sequential game that is played by two players: the learner and the opponent. At each round $t = 1, 2, \ldots, T$, the learner selects action $i(t)\ 2\ [N]$, and at the same time the opponent selects an outcome based on the opponent's strategy $p\ \ 2\ \mathsf{P}_M$, where $\mathsf{P}_n = \mathsf{f}p\ 2\ \mathsf{R}^n : p_k\ \ 0,\ \ \sum_{k=1}^{n} p_k = 1\mathsf{g}$ is the $(n\ \ 1)$-dimensional probability simplex. The outcome $j(t)$ of each round is an independent and identically distributed sample from p , and then, the learner suffers loss $\ell_{i(t);j(t)}$ at time t. The learner cannot directly observe the value of this loss, but instead observes the *feedback symbol* $y(t) = h_{i(t);j(t)}\ 2\ [A]$. The setting explained above has been widely studied in the literature of stochastic PM (Bartók et al., 2011; Komiyama et al., 2015), and we call this the *discrete* setting. In Section 4, we also introduce a *linear* setting for theoretical analysis, which is a slightly different setting from the discrete one.

The learner aims to minimize the cumulative loss over T rounds. The expected loss of action i is given by $L_i^\top p$, where L_i is the i-th column of $L^\top$. We say action i is *optimal* under strategy p if $(L_i - L_j)^\top p \leq 0$ for any $j \neq i$. We assume that the optimal action is unique, and without loss of generality that the optimal action is action 1. Let $\Delta_i = (L_i - L_1)^\top p \geq 0$ for $i \in [N]$ and $N_i(t)$ be the number of times action i is selected before the t-th round. When the time step is clear from the context, we use n_i instead of $N_i(t)$. We adopt the pseudo-regret to measure the performance: $\mathrm{Reg}(T) = \sum_{t=1}^{T} \Delta_{i(t)} = \sum_{i \in [N]} \Delta_i N_i(T+1)$. This is the relative performance of the algorithm against the *oracle*, which knows the optimal action 1 before the game starts.

We introduce the following definitions to clarify the class of PM games, for which we develop an algorithm and derive a regret upper bound. The following cell decomposition is the concept to divide the simplex Δ_M based on the loss matrix to identify the optimal action, which depends on the opponent's strategy p .

Definition 1 (Cell decomposition and Pareto-optimality (Bartók et al., 2011))**.** For every action $i \in [N]$, *cell* $C_i := \{p \in \Delta_M : (L_i - L_j)^\top p \leq 0, \forall j \neq i\}$ is the set of opponent's strategies for which action i is optimal. Action i is *Pareto-optimal* if there exists an opponent's strategy p under which action i is optimal.

Each cell is a convex closed polytope. Next, we define *neighbors* between two Pareto-optimal actions, which intuitively means that the two actions "touch" each other in their surfaces.

Definition 2 (Neighbors and neighborhood action (Bartók et al., 2011))**.** Two Pareto-optimal actions i and j are *neighbors* if $C_i \cap C_j$ is an $(M-2)$-dimensional polytope. For two neighboring actions $i, j \in [N]$, the *neighborhood action set* is defined as $N_{i,j}^+ = \{k \in [N] : C_i \cap C_j \subseteq C_k\}$.

Note that the neighborhood action set $N_{i,j}^+$ includes actions i and j from its definition. Next, we define the *signal matrix*, which encodes the information of the feedback matrix H so that we can utilize the feedback information.

Definition 3 (Signal matrix (Komiyama et al., 2015))**.** The signal matrix $S_i \in \{0,1\}^{A \times M}$ of action i is defined as $(S_i)_{y,j} = [h_{i,j} = y]$, where $[X] = 1$ if the event X is true and 0 otherwise.

Note that if we define the signal matrix as above, $S_i p \in \mathbb{R}^A$ is a probability vector over feedback symbols of action i. The following *local observability* condition separates easy and hard games, this condition intuitively means that the information obtained by taking actions in the neighborhood action set $N_{i,j}^+$ is sufficient to distinguish the loss difference between actions i and j.

Definition 4 (Local observability (Bartók et al., 2011))**.** A partial monitoring game is said to be *locally observable* if for all pairs i, j of neighboring actions, $L_i - L_j \in \sum_{k \in N_{i,j}^+} \mathrm{Im}\, S_k^\top$, where $\mathrm{Im}\, V$ is the image of the linear map V, and $V \oplus W$ is the direct sum between the vector spaces V and W.

We also consider the concept of the *strong local observability* condition, which implies the above local observability condition.

Definition 5 (Strong local observability)**.** A partial monitoring game is said to be *strongly locally observable* if for all pairs $i, j \in [N]$, $L_i - L_j \in \mathrm{Im}\, S_i^\top \cap \mathrm{Im}\, S_j^\top$.

This condition was assumed in the theoretical analysis in Vanchinathan et al. (2014), and we also assume this condition in theoretical analysis in Section 4. Note that the strong local observability means that, for any $j \neq k$, there exists $z_{j;k} \neq 0 \in \mathbb{R}^{2A}$ such that $L_j - L_k = (S_j^\top, S_k^\top)\, z_{j;k}$.

Notation. Let $\|\cdot\|$ and $\|\cdot\|_p$ be the Euclidian norm and p-norm, and let $\|x\|_A = \sqrt{x^\top A x}$ be the norm induced by the positive semidefinite matrix $A \geq 0$. Let $D_{\mathrm{KL}}(p\|q) = \sum_{a=1}^{A} p_a \log(p_a/q_a)$ be the Kullback-Leibler divergence of p from q. The vector $e_y \in \mathbb{R}^M$ is the y-th orthonormal basis of $\mathbb{R}^M$, and $\mathbf{1}_n = [1, \ldots, 1]^\top$ is the n-dimensional all-one vector. Let $q_i^{(t)}$ be the empirical feedback distribution of action i at time t, *i.e.*, $q_i^{(t)} = [n_{i1}/n_i, \ldots, n_{iA}/n_i]^\top \in \Delta_A$, where $n_{iy} = \sum_{s=1}^{t} [i(s) = i, y(s) = y]$ and $n_i = \sum_{y=1}^{A} n_{iy}$. The notation is summarized in Appendix A.

Methods for Sampling from Posterior Distribution. We briefly review the methods to draw a sample from the posterior distribution. While TS is one of the most promising algorithms, the posterior distribution can be in a quite complicated form, which makes obtaining a sample from it

Algorithm 1: TSPM Algorithm

Input: prior parameter $\lambda > 0$

Set $B_0 \quad \lambda I_M, b_0 \quad 0$.

Take each action for $n \quad 1$ times.

for $t = 1, 2, \ldots, T$ **do**

  Sample $\tilde{p}_t \quad \pi(p \mid \mathcal{H}_{s=1}^{t} \mathbb{1}[i(s), y(s)])$ based on the accept-reject sampling (Algorithm 2).

  Take action $i(t) = \arg\max_{i \in [N]} L_i^{\top} \tilde{p}_t$ and observe feedback $y(t)$.

  Update $B_t \quad B_{t-1} + S_{i(t)}^{\top} S_{i(t)}$, $b_t \quad b_{t-1} + S_{i(t)}^{\top} e_{y(t)}$.

Algorithm 2: Accept-Reject Sampling

Input: constant $R \in [0, 1]$

while true **do**

  Sample $\tilde{p}_t \quad g_t(p)$ (Algorithm 3).

  Sample $\tilde{u} \quad \mathrm{U}([0, 1])$.

  if $R\tilde{u} < F_t(\tilde{p}_t)/G_t(\tilde{p}_t)$ **then**

    return $\tilde{p}_t$.

Algorithm 3: Sampling from $g_t(p)$

Compute $\tilde{B}_t, \tilde{b}_t$ from B_t, b_t.

repeat

  Sample $p^{(\cdot)} \quad \mathrm{N}(\tilde{B}_t^{-1}\tilde{b}_t, \tilde{B}_t^{-1})$.

until $p^{(\cdot)} \in \mathcal{P}_M$;

return $\tilde{p} = [p^{(\cdot)\top}, 1 - \sum_{i=1}^{M-1}(p^{(\cdot)})_i]^{\top}$.

computationally hard. To overcome this issue, a variety of approximate posterior sampling methods have been considered, such as Gibbs sampling, Langevin Monte Carlo, Laplace approximation, and the bootstrap (Russo et al., 2018, Section 5). Recent work (Lu and Van Roy, 2017) proposed a flexible approximation method, which can even efficiently be applied to quite complex models such as neural networks. However, more recent work revealed that algorithms based on such an approximation procedure *can* suffer a linear regret (Phan et al., 2019), even if the approximation error in terms of the α-divergence is small enough.

Although BPM-TS is one of the best methods for stochastic PM, it approximates the posterior by a Gaussian distribution in a heuristic way, which can degrade the empirical performance due to the distributional discrepancy from the exact posterior distribution. Furthermore, no theoretical guarantee is provided for BPM-TS. In this paper, we mitigate these problems by providing a new algorithm for stochastic PM, which allows us to exactly draw samples from the posterior distribution. We also give theoretical analysis for the proposed algorithm.

3 Thompson-sampling-based Algorithm for Partial Monitoring

In this section, we present a new algorithm for stochastic PM games, where we name the algorithm TSPM (TS-based algorithm for PM). The algorithm is given in Algorithm 1, and we will explain the subroutines in the following.

3.1 Accept-Reject Sampling

We adopt the accept-reject sampling (Casella et al., 2004) to *exactly* draw samples from the posterior distribution. The accept-reject sampling is a technique to draw samples from a specific distribution f, and a key feature is to use a *proposal distribution* g, from which we can easily draw a sample and whose ratio to f, that is f/g, is bounded by a constant value R. To obtain samples from f, (i) we generate samples $X \quad g$; (ii) accept X with probability $f(X)/Rg(X)$. Note that f and g do not have to be normalized when the acceptance probability is calculated.

Let $\pi(p)$ be a prior distribution for p. Then an unnormalized density of the posterior distribution for p can be expressed as

$$F_t(p) = \pi(p) \prod_{i=1}^{N} \exp\left(n_i \mathrm{D}_{\mathrm{KL}}\left(q_i^{(t)} \,\middle\|\, S_i p \right) \right), \tag{1}$$

the detailed derivation of which is given in Appendix B. We use the proposal distribution with unnormalized density

$$G_t(p) = \pi(p) \prod_{i=1}^{K} \exp\left(\frac{1}{2} n_i k q_i^{(t)} \right) S_i p k^2 . \tag{2}$$

Based on these distributions, we use Algorithm 2 for exact sampling from the posterior distribution, where $\mathsf{U}([0,1])$ is the uniform distribution over $[0,1]$ and $g_t(p)$ is the distribution corresponding to the unnormalized density $G_t(p)$ in (2). The following proposition shows that setting $R = 1$ realizes the exact sampling.

Proposition 1. *Let $f_t(p)$ be the distribution corresponding to the unnormalized density $F_t(p)$ in (1). Then, the output of Algorithm 2 with $R = 1$ follows $f_t(p)$.*

This proposition can easily be proved by Pinsker's inequality, which is detailed in Appendix B.

In practice, $R \in [0,1]$ is a parameter to balance the amount of over-exploration and the computational efficiency. As R decreases from 1, the algorithm tends to accept a point p far from the mode. The case $R = 0$ corresponds the TSPM algorithm where the proposal distribution is used without the accept-reject sampling, which we call *TSPM-Gaussian*. As we will see in Section 4, TSPM-Gaussian corresponds to exact sampling of the posterior distribution when the feedback follows a Gaussian distribution rather than a multinomial distribution.

TSPM-Gaussian can be related to BPM-TS (Vanchinathan et al., 2014) in the sense that both of them use samples from Gaussian distributions. Nevertheless, they use different Gaussians and TSPM-Gaussian performs much better than BPM-TS as we will see in the experiments. Details on the relation between TSPM-Gaussian and BPM-TS are described in Appendix D.

In general, we can realize efficient sampling with a small number of rejections if the proposal distribution and the target distribution are close to each other. On the other hand, in our problem, the densities in (1) and (2) for each fixed point p exponentially decay with the number of samples n_i if the empirical feedback distribution $q_i^{(t)}$ converges. This means that $F_t(p)$ and $G_t(p)$ have an exponentially large relative gap in most rounds. Nevertheless, the number of rejections does not increase with t as we will see in the experiments, which suggests that the proposal distribution approximates the target distribution well with high probability.

3.2 Sampling from Proposal Distribution

When we consider Gaussian density $\mathsf{N}(0, \lambda I_M)$ truncated over P_M as a prior, the proposal distribution also has the Gaussian density $\mathsf{N}(B_t^{-1} b_t, B_t^{-1})$ over P_M, where

$$B_t = \lambda I_M + \sum_{i=1}^{K} n_i S_i^\top S_i = B_{t-1} + S_{i(t)}^\top S_{i(t)} , \quad b_t = \sum_{i=1}^{K} n_i S_i^\top q_i^{(t)} = b_{t-1} + S_{i(t)}^\top e_{y(t)} . \tag{3}$$

Here note that the probability simplex P_M is in an $(M-1)$-dimensional space and a sample from $\mathsf{N}(0, \lambda I_M)$ is not contained in P_M with probability one. In the literature, *e.g.,* Altmann et al. (2014), sampling methods for Gaussian distributions truncated on a simplex have been discussed. We use one of these procedures summarized in Algorithm 3, where we first sample $M-1$ elements of p from another Gaussian distribution and determine the remaining element by the constraint $\sum_{i=1}^{M} p_i = 1$.

Proposition 2. *Sampling from $g_t(p)$ is equivalent to Algorithm 3 with*

$$\tilde{B}_t = C_t - 2D_t + f_t \mathbf{1}_{M-1} \mathbf{1}_{M-1}^\top , \quad \tilde{b}_t = f_t \mathbf{1}_{M-1} - d_t + b_t^{(1:M-1)} - b_t^{(M)} \mathbf{1}_{M-1} ,$$

where $B_t = \begin{pmatrix} C_t & d_t \\ d_t^\top & f_t \end{pmatrix}$ *for* $C_t \in \mathsf{R}^{M-1 \times M-1}$, $d_t \in \mathsf{R}^{M-1}$, $f_t \in \mathsf{R}$, $b_t = [b_t^{(1:M-1)\top}, b_t^{(M)}]^\top \in \mathsf{R}^{M-1}$ *and* $D_t = \frac{1}{2}(d_t \mathbf{1}_{M-1}^\top + \mathbf{1}_{M-1} d_t^\top)$.

We give the proof of this proposition for self-containedness in Appendix C.

4 Theoretical Analysis

This section considers a regret upper bound of the TSPM algorithm.

In the theoretical analysis, we consider a *linear* setting of PM. In the linear PM, the learner suffers the expected loss $L_{i(t)}^{\top} p$ as in the discrete setting, and receives feedback vector $y(t) = S_i p + \epsilon_t$ for $\epsilon_t \sim \mathsf{N}(0, I_M)$ whereas the one-hot representation of $y(t)$ is distributed by the probability vector $S_i p$ in the discrete setting. Therefore, if ϵ_t can be regarded as a sub-Gaussian random variable as in Kirschner et al. (2020) then the linear PM includes the discrete PM, though our theoretical analysis requires ϵ_t to be Gaussian. The relation between discrete and linear settings can also be seen from the observation that bandit problems with Bernoulli and Gaussian rewards can be expressed as discrete and linear PM, respectively. The linear PM also includes the linear bandit problem, where the feedback vector is expressed as $L_i^{\top} p + \epsilon_t$.

In the linear PM, $G_t(p)$ in (2) becomes the exact posterior distribution rather than a proposal distribution. The definition of the cell decomposition for this setting is largely the same as that of discrete setting and detailed in Appendix F. Therefore, TS with exact posterior sampling in the linear PM corresponds to TSPM-Gaussian. In the linear PM, the unknown parameter p is in R^M rather than in P_M, and therefore we consider the prior $\pi(p) = \mathsf{N}(0, \lambda I_M)$ over R^M, where the posterior distribution becomes $\mathsf{N}(B_t^{-1} b_t, B_t^{-1})$.

There are a few works that analyze TS for the PM because of its difficulty. For example in Vanchinathan et al. (2014), an analysis of the TS-based algorithm (BPM-TS) is not given despite the fact that its performance is better than the algorithm based on a confidence ellipsoid (BPM-LEAST). Zimmert and Lattimore (2019) considered the theoretical aspect of a variant of TS for the linear PM in view of the Bayes regret, but this algorithm is based on the knowledge on the time horizon and different from the family of TS used in practice. More specifically, 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.

Types of Regret Bounds. We focus on the *(a) problem-dependent (b) expected pseudo-regret*. (a) In the literature, a *minimax* (or *problem-independent*) regret bound has mainly been considered, for example, to classify difficulties of the PM problem (Bartók et al., 2010; Bartók et al., 2011). On the other hand, a *problem-dependent* regret bound often reflects the empirical performance more clearly than the minimax regret (Bartók et al., 2012; Vanchinathan et al., 2014; Komiyama et al., 2015). For this reason, we consider this problem-dependent regret bound. (b) In complicated settings of bandit problems, a *high-probability regret bound* has mainly been considered (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013b), which bounds the pseudo-regret with high probability $1 - \delta$. Though such a bound can be transformed to an expected regret bound, this type of analysis often sacrifices the tightness since a linear regret might be suffered with small probability δ. This is why the analysis in Vanchinathan et al. (2014) for BPM-LEAST finally yielded an $\tilde{\mathrm{O}}(\sqrt{T})$ expected regret bound whereas their high-probability bound is $\mathrm{O}(\log T)$.

4.1 Regret Upper Bound

In the following theorem, we show that logarithmic problem-dependent expected regret is achievable by the TSPM-Gaussian algorithm.

Theorem 3 (Regret upper bound)**.** *Consider any finite stochastic linear partial monitoring game. Assume that the game is strongly locally observable and* $\Delta_i = (L_i - L_1)^{\top} p > 0$ *for any* $i \neq 1$. *Then, the regret of TSPM-Gaussian satisfies for sufficiently large* T *that*

$$\mathsf{E}[\mathrm{Reg}(T)] = \mathrm{O}\left(\frac{AN^2 M \max_{i \in [N]} \Delta_i}{\Lambda^2} \log T\right), \tag{4}$$

where $\Lambda := \min_{i \neq 1} \Lambda_i$ *for* $\Lambda_i = \Delta_i / \|z_{1,i}\|$ *with* $z_{1,i}$ *defined after Definition 5.*

Remark. In the proof of Theorem 3, it is sufficient to assume that $L_1 - L_i \in \mathrm{Im}\, S_1^{\top} - \mathrm{Im}\, S_i^{\top}$ for $i \in [N]$, which is weaker than the strong local observability, though it is still sometimes stronger than the local observability condition.

The proof of Theorem 3 is given in Appendix F. This result is the first problem-dependent bound of TS for PM, which also becomes the first logarithmic regret bound of TS for linear bandits.

The norm of $z_{j,k}$ in Λ intuitively 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 $z_{j,k}$.

Unfortunately, the derived bound in Theorem 3 has quadratic dependence on N, which seems to be not tight. This quadratic dependence comes from the difficulty of the *expected* regret analysis. In general, we evaluate the regret before and after the convergence of the statistics separately. Whereas the latter one usually becomes dominant, the main difficulty comes from the analysis of the former one, which might become large with low probability (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013a).

In our analysis, we were not able to bound the former one within a non-dominant order, though it is still logarithmic in T. In fact, our analysis shows that the regret after convergence is $\mathrm{O}(\sum_{i \neq 1} \Delta_i \frac{A}{2} \log T)$ as shown in Lemma 18 in Appendix F, which will become the regret with high probability. In particular, if we consider the classic bandit problem as a PM game, we can confirm that the derived bound after convergence becomes the best possible bound

$$\mathrm{O}\left(\sum_{i \neq 1} \frac{\log T}{\Delta_i}\right)$$

by considering Λ_i depending on each suboptimal arm i as the difficulty measure instead of Λ. Still, deriving a regret bound for the term before convergence within an non-dominant order is an important future work.

4.2 Technical Difficulties of the Analysis

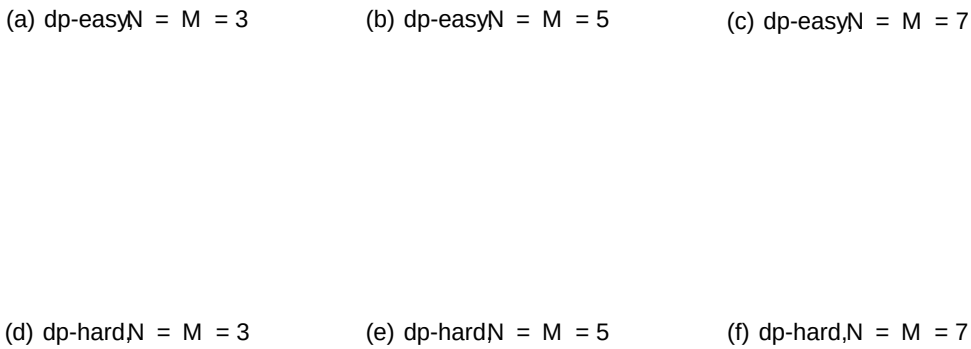
The main difficulty of this regret analysis is that PM requires to consider the statistics of *all* actions when the number of selections $N_i(t)$ of some action i is evaluated. This is in stark contrast to the analysis of the classic bandit problems, where it becomes sufficient to evaluate statistics of action i and the best action 1. This makes the analysis remarkably complicated in TS, where we need to separately consider the randomness caused by the feedback and TS.

To overcome this difficulty, we handle the effect of actions of no interest in two different novel ways depending on each decomposed regret. The first one is to evaluate the worst-case effect of these actions based on an argument (Lemma 10) related to the law of the iterated logarithm (LIL), which is sometimes used in the best-arm identification literature to improve the performance (Jamieson et al., 2014). The second one is to bound the action-selection probability of TS using an argument of (super-)martingale (Theorem 16), which is of independent interest. Whereas such a technique is often used for the construction of confidence bounds (Abbasi-Yadkori et al., 2011), we reveal that it is also useful for evaluation of the regret of TS.

We only focused on the Gaussian noise $\epsilon_t \sim \mathrm{N}(0, I_M)$, rather than the more general sub-Gaussian noise. This 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. 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 of the problem-dependent regret analysis, 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.

5 Experiments

In this section, we numerically compare the performance of TSPM and TSPM-Gaussian against existing methods, which are RandomPM (the algorithm which selects action randomly), FeedExp3 (Piccolboni and Schindelhauer, 2001), and BPM-TS (Vanchinathan et al., 2014). Recently, Lattimore and Szepesvári (2019) considered the sampling-based algorithm called Mario sampling for easy games. Mario sampling coincides with TS (except for the difference between pseudo-regret and regret with known time horizon) mentioned in the last section when any pair of actions is a neighbor. As shown in Appendix G, this property is indeed satisfied for dp-easy games defined in the following. Therefore, the performance is essentially the same between TSPM with $R = 1$ and Mario sampling. To compare the performance, we consider a dynamic pricing problem, which is a typical example of PM games. We conducted experiments on the discrete setting because the experiments for PM has been mainly focused on the discrete setting.



(a) dp-easy, $N = M = 3$ (b) dp-easy, $N = M = 5$ (c) dp-easy, $N = M = 7$

(d) dp-hard, $N = M = 3$ (e) dp-hard, $N = M = 5$ (f) dp-hard, $N = M = 7$

Figure 1: Regret-round plots of algorithms. The solid lines indicate the average over 100 independent trials. The thin fillings are the standard error.

In the dynamic pricing game, the player corresponds to a seller, and the opponent corresponds to a buyer. At each round, the seller sells an item for a specific price $i(t)$, and the buyer comes with an evaluation price $j(t)$ for the item, where the selling price and the evaluation price correspond to the action and outcome, respectively. The buyer buys the item if the selling price $i(t)$ is smaller than or equal to $j(t)$ and not otherwise. The seller can only know if the buyer bought the item (denoted as feedback 0) or did not buy the item (denoted as 1). The seller aims to minimize the cumulative loss, and there are two types of definitions for the loss, where each induced game falls into the easy and hard games. We call them dp-easy and dp-hard games, respectively.

In both cases, the seller incurs the constant loss $c > 0$ when the item is not bought due to the loss of opportunity to sell the item. In contrast, when the item is not bought, the loss incurred to the seller is different between these settings. The seller in the dp-easy game does not take the buyer's evaluation price into account. In other words, the seller gains the selling price $i(t)$ as a reward (equivalently incurs $i(t)$ as a loss). Therefore, the loss for the selling price $i(t)$ and the evaluation $j(t)$ is

$$\ell_{i(t),j(t)} = i(t)\mathbf{1}[i(t) \le j(t)] + c\mathbf{1}[i(t) > j(t)].$$

This setting can be regarded as a generalized version of the online posted price mechanism, which was addressed in, e.g., Blum et al. (2004) and Cesa-Bianchi et al. (2006), and an example of strongly locally observable games.

On the other hand, the seller in dp-hard game does take the buyer's evaluation price into account when the item is bought. In other words, the seller incurs the difference between the opponent evaluation and the selling price $j(t) - i(t)$ as a loss because the seller could have made more profit if the seller had sold at the price $j(t)$. Therefore, the loss incurred at time is

$$\ell_{i(t),j(t)} = (j(t) - i(t))\mathbf{1}[i(t) \le j(t)] + c\mathbf{1}[i(t) > j(t)].$$

This setting is also addressed in Cesa-Bianchi et al. (2006), and belongs to the class of hard games. Note that our algorithm can also be applied to a hard game, though there is no theoretical guarantee.

Setup. In the both dp-easy and dp-hard games, we fixed $M \in \{3, 5, 7\}$ and $c = 2$. We fixed the time horizon T to 10000 and simulated 100 times. For FeedExp3 and BPM-TS, the setup of hyperparameters follows their original papers. For TSPM, we set $\varepsilon = 0.001$, and R was selected from $\{0.01, 1.0\}$. Here, recall that TSPM with $R = 1$ and $R = 0$ correspond to the exact sampling and TSPM-Gaussian, respectively, and a smaller value of R gives the higher acceptance probability in the accept-reject sampling. Therefore, using small R makes the algorithm time-efficient, although it can worsen the performance since it over-explores the tail of the posterior distributions. To stabilize

(a) dp-easy, $N = M = 3$ (b) dp-easy, $N = M = 5$ (c) dp-easy, $N = M = 7$

Figure 2: The number of rejected times by the accept-reject sampling. The solid lines indicate the average over 100 independent trials after taking moving average with window size 100.

sampling from the proposal distribution in Algorithm 3, we used an initialization that takes each action $n = 10A$ times. The detailed settings of the experiments with more results are given in Appendix H.

Results. Figure 1 is the empirical comparison of the proposed algorithms against the benchmark methods. This result shows that, in all cases, the TSPM with exact sampling gives the best performance. TSPM-Gaussian also outperforms BPM-TS even though both of them use Gaussian distributions as posteriors. Besides, the experimental results suggest that our algorithm performs reasonably well even for a hard game. It can be observed that the proposed methods outperform BPM-TS more significantly for a larger number of outcomes. Further discussion for this observation is given in Appendix D.

Figure 2 shows the number of rejections at each time step in the accept-reject sampling. We counted the number of times that either Line 4 in Algorithm 2 or Line 4 in Algorithm 3 was not satisfied. In the accept-reject sampling, it is desirable that the frequency of rejection does not increase as the time-step t and does not increase rapidly with the number of outcomes. We can see that the former one is indeed satisfied. For the latter property, the frequency of rejection becomes unfortunately large when exact sampling ($R = 1$) is conducted. 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.

6 Conclusion and Discussion

This paper investigated Thompson sampling (TS) for stochastic partial monitoring from the algorithmic and theoretical viewpoints. We provided a new algorithm that enables exact sampling from the posterior distribution, and numerically showed that the proposed algorithm outperforms existing methods. Besides, we provided an upper bound for the problem-dependent logarithmic expected pseudo-regret for the linearized version of the partial monitoring. To our knowledge, this bound is the first logarithmic problem-dependent expected pseudo-regret bound of a TS-based algorithm for linear bandit problems and strongly locally observable partial monitoring games.

There are several remaining questions. As mentioned in Section 4, Kirschner et al. (2020) considered linear partial monitoring with the feedback structure $y(t) = S_{i(t)}p + \epsilon_t$, where $(\epsilon_t)_{t=1}^T$ is a sequence of independent sub-Gaussian noise vectors in $\mathbb{R}^M$. This setting is the generalization of our linear setting, where $(\epsilon_t)_{t=1}^T$ are i.i.d. Gaussian vectors. Therefore, a natural question that arises is whether we can extend our analysis on TSPM-Gaussian to the sub-Gaussian case, although we believe it would be not straightforward as discussed in Section 4. It is also an important open problem to derive a regret bound on TSPM using the exact posterior sampling for the discrete partial monitoring. Although we conjecture that the algorithm also achieves logarithmic regret for the setting, there still remain some difficulties in the analysis. In particular, we have to handle the KL divergence $f_{i,t}(p)$ and consider the restriction of the support of the opponent's strategy to $\mathcal{P}_M$, which make the analysis much more complicated. Besides, it is worth noting that the theoretical analysis of TS for hard games has never been theoretically investigated. We believe that in general TS suffers linear regret in the minimax sense due to its greediness. However, we conjecture that TS can achieve the sub-linear regret for some specific instances of hard games in the sense of the problem-dependent regret, as empirically observed in the experiments. Finally, it is an important open problem to derive the minimax regret for anytime TS-based algorithms. This needs more detailed analysis of $o(\log T)$ terms in the regret bound, which were dropped in our main result.

Broader Impact

Application. Partial monitoring (PM) includes various online decision-making problems such as multi-armed bandits, linear bandits, dynamic pricing, and label e cient prediction. Not only can PM handles them, the dueling bandits, combinatorial bandits, transductive bandits, and many other problems can be seen as a partial monitoring game, as discussed in Kirschner et al. (2020). Therefore, our analysis of Thompson sampling (TS) for PM games pushes the application of TS to a more wide range of online decision-making problems forward. Moreover, PM has the potential that novel online-decision making problems are newly discovered, where we have to handle the limited feedback in an online fashion.

Practical Use. The obvious advantage of using TS is that the users can easily apply the algorithm to their problems. They do not have to solve mathematical optimization problems, which are often required to solve when using non-sampling-based algorithms (Bartók et al., 2012; Komiyama et al., 2015). For the negative side, the theoretical analysis for the regret upper bound might make the users become overcon dent when the users use their algorithms. For example, they might use the TSPM algorithm to the linear PM game with heavy-tailed noise, such as sub-exponential noise, without noticing it. Nevertheless, this is not an TS-speci c problem, but one that can be found in many theoretical studies, and TS is still one of the most promising policies.

Acknowledgements

The authors would like to thank the meta-reviewer and reviewers for a lot of helpful comments. The authors would like to thank Kento Nozawa and Ikko Yamane for maintaining servers for our experiments, and Kenny Song for helpful discussion on the writing. TT was supported by Toyota-Dwango AI Scholarship, and RIKEN Junior Research Associate Program for the nal part of the project. JH was supported by KAKENHI 18K17998, and MS was supported by KAKENHI 17H00757.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312 2320, 2011.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *the 25th Annual Conference on Learning Theory*, volume 23, pages 39.1 39.26, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *the Sixteenth International Conference on Arti cial Intelligence and Statistics*, volume 31, pages 99 107, 2013a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payo s. In *the 30th International Conference on Machine Learning*, pages 127 135, 2013b.

Yoann Altmann, Steve McLaughlin, and Nicolas Dobigeon. Sampling from a multivariate gaussian distribution truncated on a simplex: A review. In *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pages 113 116, 2014.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classi cation of nite partial-monitoring games. In *Algorithmic Learning Theory*, pages 224 238, 2010.

Gábor Bartók, Navid Zolghadr, and Csaba Szepesvári. An adaptive algorithm for nite stochastic partial monitoring. In *the 29th International Conference on Machine Learning*, pages 1 20, 2012.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of nite partial-monitoring games in stochastic environments. In *the 24th Annual Conference on Learning Theory*, volume 19, pages 133 154, 2011.

Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. *Theoretical Computer Science*, 324(2):137 146, 2004.

George Casella, Christian P. Robert, and Martin T. Wells. Generalized accept-reject sampling schemes volume 45 of *Lecture Notes Monograph Series*, pages 342 347. Institute of Mathematical Statistics, 2004.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label e cient prediction. *IEEE Transactions on Information Theory*, 51(6):2152 2162, 2005.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. Mathematics of Operations Research, 31(3):562 580, 2006.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In Advances in Neural Information Processing Systems, pages 2249 2257, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In 21st Annual Conference on Learning Theory, pages 355 366, 2008.

Junya Honda and Akimichi Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. In the Seventeenth International Conference on Arti cial Intelligence and Statistics, volume 33, pages 375 383, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. The 27th Conference on Learning Theory, pages 423 439, 2014.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal nite-time analysis. In Algorithmic Learning Theory, pages 199 213, 2012.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. arXiv preprint arXiv:2002.11182, 2020.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In the 44th Annual IEEE Symposium on Foundations of Computer Science, pages 594 605, 2003.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in nite stochastic partial monitoring. In Advances in Neural Information Processing Systems 28, pages 1792 1800, 2015.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In Advances in Neural Information Processing Systems, pages 1448 1456, 2013.

T. L. Lai and Herbert Robbins. Asymptotically e cient adaptive allocation rules. Advances in Applied Mathematics, 6(1):4 22, 1985.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In the 32nd Annual Conference on Learning Theory, volume 99, pages 2111 2139, 2019.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. Advances in Neural Information Processing Systems, 30, pages 3258 3266, 2017.

My Phan, Yasin Abbasi Yadkori, and Justin Domke. Thompson sampling and approximate inference. In Advances in Neural Information Processing Systems, pages 8804 8813, 2019.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In COLT/EuroCOLT, pages 208 223, 2001.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. Foundations and Trends in Machine Learning, 11(1):1 96, 2018.

Aldo Rustichini. Minimizing regret: The general case. Games and Economic Behavior, 29(1): 224 243, 1999.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3-4):285 294, 12 1933.

Hastagiri P Vanchinathan, Gábor Bartók, and Andreas Krause. E cient partial monitoring with prior information. In Advances in Neural Information Processing Systems, pages 1691 1699, 2014.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, thompson sampling and the information ratio. In Advances in Neural Information Processing Systems, pages 11973 11982, 2019.

Martin Zinkevich. Online convex programming and generalized in nitesimal gradient ascent. In the Twentieth International Conference on Machine Learning, pages 928 935. AAAI Press, 2003.

A Notation

Table 1 summarizes the symbols used in this paper.

Table 1: List of symbols used in this paper.

Symbol	Meaning
P_n	$(n-1)$-dimensional probability simplex
$\|\cdot\|$	Euclidian norm for vector and operator norm for matrix
$\|\cdot\|_p$	p-norm
$\|\cdot\|_A$	norm induced by positive semide nite matrix A
$D_{KL}(p\|q)$	KL divergence from q to p
$B_r^n(p)$	n-dimensional Euclidian ball of radius r at point $p \in \mathbb{R}^N$
$N, M \in \mathbb{N}$	the number of actions and outcomes
	set of feedback symbols
A	the number of feedback symbols
$p \in P_M$	opponent's strategy
T	time horizon
$L = (\ell_{i,j}) \in \mathbb{R}^{N \times M}$	loss matrix
$H = (h_{i,j}) \in$	$^{N \times M}$ feedback matrix
$S_i \in \{0,1\}^{A \times M}$ $(i = 1, \dots, N)$	signal matrix
$i(t)$	action taken at time t
$N_i(t)$	the number of times the action i is taken before time $t \in [T]$
$j(t)$	outcome taken by opponent at time t
$y(t)$	feedback observed at time t
$F_t(p)$	unnormalized posterior distribution in (1)
$f_t(p)$	probability density function corresponding to $F_t(p)$
$G_t(p)$	unnormalized proposal distribution for $F_t(p)$ in (2)
$g_t(p)$	probability density function corresponding to $G_t(p)$
$q_i^{(t)} \in P_M$	empirical feedback distribution of action i by time t
$q_{i;n} \in P_M$	empirical feedback distribution of action i after the action is taken n times
$C_i \subset P_M$	cell of action i

B Posterior Distribution and Proposal Distribution in Section 3

In this appendix, we discuss representation of the posterior distribution and its relation with the proposal distribution.

Proposition 4. $F_t(p)$ in (1) is proportional to the posterior distribution of the opponent's strategy, and $F_t(p) \geq G_t(p)$ for all $p \in P_M$.

Proof. The posterior distribution of the opponent's strategy parameter for $\{i(s), y(s)\}_{s=1}^t$ is rewritten as

$$\begin{aligned} p\left(\{i(s), y(s)\}_{s=1}^t\right) &\propto p\left(\{i(s), y(s)\}_{s=1}^t\right) \\ &\propto \varphi(p) \prod_{s=1}^t P\{y(s) \mid i(s), p\} \\ &= \varphi(p) \prod_{i=1}^N \prod_{y=1}^A (S_{i,y}\, p)^{n_{iy}} \\ &\propto \varphi(p) \prod_{i=1}^N \exp\left\{ n_i D_{KL}\left(q_i^{(t)} \,\middle\|\, S_i p\right) \right\}; \end{aligned} \tag{5}$$

where $S_{i,y}$ is the i-th row of the signal matrix S_i, and note that $q_i^{(t)}$ is the empirical feedback distribution of action i at time t, that is, $q_i^{(t)} = [n_{i1}/n_i, \dots, n_{iA}/n_i]^\top \in P_A$ for $n_{iy} = \sum_{s=1}^t \mathbb{1}[i(s) = i, y(s) = y]$ and $n_i = \sum_{y=1}^A n_{iy}$.

Next, we show that $F_t(p) \le G_t(p)$ holds for all $p \in P_M$. Using the Pinsker's inequality, the unnormalized posterior distribution $F_t(p)$ can be bounded from above as

$$\begin{aligned}
F_t(p) &= \pi(p) \prod_{i=1}^{N} \exp\left\{ n_i D_{KL}\left(q_i^{(t)} \,\|\, S_i p \right) \right\} \\
&\le \pi(p) \prod_{i=1}^{N} \exp\left\{ \frac{1}{2} n_i \| q_i^{(t)} - S_i p \|_1^2 \right\} \quad \text{(by Pinsker's ineq.)} \\
&= \pi(p) \exp\left\{ \frac{1}{2} \sum_{i=1}^{N} n_i \| q_i^{(t)} - S_i p \|_1^2 \right\} \\
&\le \pi(p) \exp\left\{ \frac{1}{2} \sum_{i=1}^{N} n_i \| q_i^{(t)} - S_i p \|^2 \right\} \quad \text{by } \| q_i^{(t)} - S_i p \|_1 \le \sqrt{|\mathcal{X}|}\, \| q_i^{(t)} - S_i p \| \\
&= G_t(p). \qquad (6)
\end{aligned}$$

□

Remark. The unnormalized density $G_t(p)$ is indeed Gaussian. Recalling that B_t and b_t are defined in (3) as

$$B_t = \sum_{i=1}^{N} n_i S_i^\top S_i = \sum_{s=1}^{t} S_{i(s)}^\top S_{i(s)} = B_{t-1} + S_{i(t)}^\top S_{i(t)}; \quad b_t = \sum_{i=1}^{N} n_i S_i^\top q_i^{(t)} = b_{t-1} + S_{i(t)}^\top e_{y(t)}; \qquad (7)$$

we have

$$\begin{aligned}
\sum_{i=1}^{N} n_i \| q_i^{(t)} - S_i p \|^2 &= \sum_{i=1}^{N} n_i (q_i^{(t)} - S_i p)^\top (q_i^{(t)} - S_i p) \\
&= p^\top \underbrace{\sum_{i=1}^{N} n_i S_i^\top S_i}_{B_t} p - 2 \underbrace{\sum_{i=1}^{N} n_i S_i^\top q_i^{(t)}}_{b_t}{}^\top p + \underbrace{\sum_{i=1}^{N} n_i \| q_i^{(t)} \|^2}_{c_t} \\
&= p^\top B_t p - 2 b_t^\top p + c_t \\
&= (p - B_t^{-1} b_t)^\top B_t (p - B_t^{-1} b_t) + c_t - b_t^\top B_t^{-1} b_t. \qquad (8)
\end{aligned}$$

Therefore, we have

$$\exp\left\{ \frac{1}{2} \sum_{i=1}^{N} n_i \| q_i^{(t)} - S_i p \|^2 \right\} \propto \exp\left\{ \frac{1}{2} (p - B_t^{-1} b_t)^\top B_t (p - B_t^{-1} b_t) \right\}. \qquad (9)$$

C Proof of Proposition 2

We will see that the the procedure sampling from $g_t(p)$ and Algorithm 3 are equivalent. First, we derive the Gaussian density $g_t(p)$ projected onto $\{p \in \mathbb{R}^M : \sum_{i=1}^{M} p_i = 1\}$.

For simplicity, we omit the subscript t and write, e.g., B instead of B_t. We define $p = [p_{(-)}^\top; p_M]^\top \in \mathbb{R}^{M-1}$. Let $h = B^{-1} b$, and define $h = [h_{(-)}^\top; h_M]^\top \in \mathbb{R}^{M-1}$. Let $B = \begin{bmatrix} C & d \\ d^\top & f \end{bmatrix}$, where $C \in \mathbb{R}^{(M-1)\times(M-1)}$, $d \in \mathbb{R}^{M-1}$, and $f \in \mathbb{R}$. Also, let $b = [b_{(-)}^\top; b^{(M)}]^\top \in \mathbb{R}^{M-1}$.

Using the decomposition

$$(p - B^{-1} b)^\top B (p - B^{-1} b) = \underbrace{p_{(-)}^\top B p_{(-)}}_{\text{(a)}} - 2 \underbrace{h_{(-)}^\top B p_{(-)}}_{\text{(b)}} + h^\top B h; \qquad (10)$$

we rewrite each term by restricting the domain of p so that it satisfies the condition $\sum_{i=1}^{M} p_i = 1$. Now the first term (a) is rewritten as

$$\begin{aligned}(a) &= p^{(-M)\top} C p^{(-M)} + 2 p^{(-M)\top} d p_M + f p_M^2 \\ &= p^{(-M)\top} C p^{(-M)} + 2 p^{(-M)\top} d \Big(1 - \sum_{i=1}^{M-1} p_i\Big) + f \Big(1 - \sum_{i=1}^{M-1} p_i\Big)^2 : \qquad (11)\end{aligned}$$

The term (a1) is rewritten as

$$\begin{aligned}(a1) &= p^{(-M)\top} d - p^{(-M)\top} d \sum_{i=1}^{M-1} p_i \\ &= p^{(-M)\top} d - p^{(-M)\top} d 1_{M-1}^\top p^{(-M)} \\ &= p^{(-M)\top} d - p^{(-M)\top} D p^{(-M)} \qquad D = \frac{1}{2}\big(d 1_{M-1}^\top + 1_{M-1} d^\top\big) ; \qquad (12)\end{aligned}$$

and the term (a2) is rewritten as

$$\begin{aligned}(a2) &= \Big(1 - \sum_{i=1}^{M-1} p_i\Big)^2 \\ &= 1 - 2\sum_{i=1}^{M-1} p_i + \Big(\sum_{i=1}^{M-1} p_i\Big)^2 \\ &= 1 - 2 1_{M-1}^\top p^{(-M)} + p^{(-M)\top} 1_{M-1} 1_{M-1}^\top p^{(-M)} : \qquad (13)\end{aligned}$$

Therefore,

$$(a) = p^{(-M)\top} \underbrace{\big(C - 2D + f\, 1_{M-1} 1_{M-1}^\top\big)}_{B} p^{(-M)} - 2(f 1_{M-1} - d)^\top p^{(-M)} + f : \qquad (14)$$

With regard to the term (b), we have

$$\begin{aligned}(b) &= b^\top p \\ &= b^{(-M)\top} p^{(-M)\top} + b^{(M)} p_M \\ &= (b^{(-M)} - b^{(M)} 1_{M-1})^\top p^{(-M)} + b^{(M)} : \qquad (15)\end{aligned}$$

Therefore,

$$\begin{aligned}&(p - B^{-1} b)^\top B (p - B^{-1} b) \\ &= p^{(-M)\top} B p^{(-M)} - 2\big(\underbrace{f 1_{M-1} - d + b^{(-M)} - b^{(M)} 1_{M-1}}_{b}\big)^\top p^{(-M)} + f - 2 b^{(M)} + h^\top B h \\ &= (p^{(-M)} - B^{-1} b)^\top B (p^{(-M)} - B^{-1} b) + f - 2 b^{(M)} - b^\top B^{-1} b + b^\top B^{-1} b \qquad \text{by } h^\top B h = b^\top B^{-1} b : \\ &\qquad (16)\end{aligned}$$

From the above argument, the density $N(B^{-1} b; B^{-1})$ is the Gaussian distribution of (p) on $\{p \in R^M : \sum_{i=1}^{M} p_i = 1\}$. Therefore, the $p = [p^{(-M)\top}; 1 - \sum_{i=1}^{M-1} (p^{(-M)})_i]^\top$ for $p^{(-M)} \sim N(B^{-1} b; B^{-1})$ is supported over $\{p \in R^M : \sum_{i=1}^{M} p_i = 1\}$.

If the sample $p^{(-M)}$ from $N(B^{-1} b; B^{-1})$ is in P_{M-1}, then we can obtain the last element $p^{(M)}$ by $p^{(M)} = 1 - \sum_{i=1}^{M-1} (p^{(-M)})_i$. Otherwise, the probability that $p^{(-M)}$ is the first $M-1$ elements of the sample from $g_t(p)$ is zero, and hence $[p^{(-M)\top}; p^{(M)}]^\top$ cannot be a sample from $g_t(p)$. Therefore, sampling p_t from $g_t(p)$ and Algorithm 3 are equivalent.

D Relation between TSPM-Gaussian and BPM-TS

In this appendix, we discuss the relation between TSPM-Gaussian and BPM-TS (Vanchinathan et al., 2014).

Underlying Feedback Structure. Here, we discuss the underlying feedback structure behind TSPM-Gaussian and BPM-TS.

We first consider the underlying feedback structure behind BPM-TS. In the following, we see that the feedback structure

$$y(t) = S_{i(t)}p + \varepsilon_{i(t)} \; ; \qquad \varepsilon \sim N(0; I_M) \tag{17}$$

induces the posterior distribution in BPM-TS. Under this feedback structure, we have $\mu(t) \sim N(S_{i(t)}p; S_{i(t)}S_{i(t)}^{\top})$.

When we take the prior distribution $\pi(p)$ as $N(0; \sigma_0^2 I_M)$, the posterior distribution for the opponent's strategy parameter can be written as

$$\begin{aligned}
&\pi\left(p \mid \{i(s), y(s)\}_{s=1}^{t}\right) \\
&\propto \pi(p) \prod_{s=1}^{t} f(y(s) \mid i(s), p) \\
&= \pi(p) \prod_{s=1}^{t} P_{y \sim N(S_{i(s)}p, S_{i(s)}S_{i(s)}^{\top})}\{y = y(s)\} \\
&= \exp\left(-\frac{p^{\top}p}{2\sigma_0^2}\right) \prod_{s=1}^{t} \exp\left(-\frac{1}{2}(y(s) - S_{i(s)}p)^{\top}(S_{i(s)}S_{i(s)}^{\top})^{-1}(y(s) - S_{i(s)}p)\right) \\
&= \exp\left(-\frac{1}{2}p^{\top}\left(\frac{1}{\sigma_0^2}I_M + \sum_{s=1}^{t} S_{i(s)}^{\top}(S_{i(s)}S_{i(s)}^{\top})^{-1}S_{i(s)}\right)p\right. \\
&\qquad \left. + 2\sum_{s=1}^{t} y(s)^{\top}(S_{i(s)}S_{i(s)}^{\top})^{-1}S_{i(s)}p + (\text{a term independent of } p)\right) \\
&\propto \exp\left(-\frac{1}{2}(p^{\top}B_t^{BPM}p - 2b_t^{BPM\top}p)\right) \\
&\propto \exp\left(-\frac{1}{2}(p - B_t^{BPM-1}b_t^{BPM})^{\top}B_t^{BPM}(p - B_t^{BPM-1}b_t^{BPM})\right) ;
\end{aligned} \tag{18}$$

where

$$B_t^{BPM} = \frac{1}{\sigma_0^2}I_M + \sum_{s=1}^{t} S_{i(s)}^{\top}(S_{i(s)}S_{i(s)}^{\top})^{-1}S_{i(s)} = B_{t-1}^{BPM} + S_{i(t)}^{\top}(S_{i(t)}S_{i(t)}^{\top})^{-1}S_{i(t)} ; \tag{19}$$

$$b_t^{BPM} = \sum_{s=1}^{t} S_{i(s)}^{\top}(S_{i(s)}S_{i(s)}^{\top})^{-1}y(s) = b_{t-1}^{BPM} + S_{i(t)}^{\top}(S_{i(t)}S_{i(t)}^{\top})^{-1}y(t) : \tag{20}$$

Therefore, the posterior distribution $p \mid \{i(s), y(s)\}_{s=1}^{t}$ is

$$q\frac{1}{(2\pi)^M |B_t^{BPM-1}|} \exp\left(-\frac{1}{2}(p - B_t^{BPM-1}b_t^{BPM})^{\top}B_t^{BPM}(p - B_t^{BPM-1}b_t^{BPM})\right) : \tag{21}$$

and this distribution indeed corresponds to the posterior distribution in BPM-TS (Vanchinathan et al., 2014) with $B_t^{BPM} = \Sigma_t^{-1}$.

Using the same argument, we can confirm that the feedback structure

$$y_t = S_i p + \varepsilon \; ; \qquad \varepsilon \sim N(0; I_M) : \tag{22}$$

induces

$$g_t(p) := \sqrt{\frac{1}{(2\pi)^M |B_t^{-1}|}} \exp\left(-\frac{1}{2}\|p - B_t^{-1} b_t\|_{B_t}^2\right); \tag{23}$$

which corresponds to the posterior distribution for TSPM in linear partial monitoring.

Covariances in TSPM-Gaussian and BPM-TS.In the linear partial monitoring, TSPM assumes noise with covariance I_M, which is compatible with the fact that the discrete setting can be regarded as linear PM with I_M-sub-Gaussian noise. On the other hand, BPM-TS assumes covariance $S_i S_i^\top$, and in general $I_M \preceq S_i S_i^\top$ holds. Therefore, BPM-TS assumes unnecessarily larger covariance, which makes learning slow down.

E Preliminaries for Regret Analysis

In this appendix, we give some technical lemmas, which are used for the derivation of the regret bound in Appendix F. Here, we write $X \preceq Y$ to denote $Y - X \succeq 0$. For $a, b \in \mathbb{R}$, let $a \wedge b$ be a if $a \le b$ otherwise b, and $a \vee b$ be b if $a \le b$ otherwise a. We use $h(a) := P_{X \sim \chi^2_M}\{X \ge a\}$ to evaluate the behavior of the posterior samples, where χ^2_M is the chi-squared distribution with M degree of freedom.

E.1 Basic Lemmas

Fact 5 (Moment generating function of squared-Gaussian distribution). Let X be the random variable following the standard normal distribution. Then, the moment generating function of X^2 is $\mathbb{E}\exp(\lambda X^2) = (1 - 2\lambda)^{-1/2}$ for $\lambda < 1/2$.

Lemma 6 (Chernoff bound for chi-squared random variable). Let X be the random variable following the chi-squared distribution with k degree of freedom. Then, for any $a > 0$ and $0 < \lambda < 1/2$,

$$P\{X \ge a\} \le e^{-\lambda a}(1 - 2\lambda)^{-\frac{k}{2}}. \tag{24}$$

Proof. By Markov's inequality, the LHS can be bounded as

$$\begin{aligned} P\{X \ge a\} &= P\left(\sum_{i=1}^k X_i^2 \ge a\right) \quad (X_1, \dots, X_k \overset{i.i.d.}{\sim} N(0,1)) \\ &= P\left(\exp\left(\lambda \sum_{i=1}^k X_i^2\right) \ge \exp(\lambda a)\right) \\ &\le e^{-\lambda a} \mathbb{E}\left[e^{\lambda X_1^2}\right]^k \quad \text{(by Markov's ineq)} \\ &= e^{-\lambda a}(1 - 2\lambda)^{-\frac{k}{2}} \quad \text{(by Fact 5)}; \end{aligned} \tag{25}$$

which completes the proof. □

E.2 Property of Strong Local Observability

Recall that $\Delta_i = (L_i - L_1)^\top p > 0$ for $i \in [N]$, which is the difference of the expected loss of actions i and 1. For this define

$$\varepsilon := \frac{1}{2\sqrt{A}} \min_{i \in [N]} \frac{\Delta_i}{\|z_{1,i}\|} \wedge \min_{p' \in C_1^c} \frac{4}{3}\|p - p'\|; \tag{26}$$

which is used throughout the proof of this appendix and Appendix F. The following lemma provides the key property of the strong local observability condition.

Lemma 7. For any partial monitoring game with strong local observability and $p \in \mathbb{R}^M$, any of the conditions 1–3 in the following is not satisfied:

1. $L_1^\top p > L_k^\top p$ (Worse action k looks better under p.)
2. $\|S_1^\top p\| \le \|S_1^\top p - \|$

3. $\|S_k p - S_k p'\| < \epsilon$.

Proof. We prove by contradiction. Assume that there exists $p, p' \in \mathbb{R}^M$ such that conditions 1–3 are simultaneously satisfied.

Now, by the conditions 2 and 3, we have

$$\begin{aligned} |S_1 p - S_1 p'| &\le 1_A, \\ |S_k p - S_k p'| &\le 1_A. \end{aligned} \tag{27}$$

Here, $|\cdot|$ is the element-wise absolute value, and $\le$ means that the inequality holds for each element. Therefore,

$$\begin{bmatrix} S_1 \\ S_k \end{bmatrix} (p - p') \le 1_{2A}. \tag{28}$$

On the other hand, by the strong local observability condition, for each $k \ne 1$, there exists $z_{1;k} \ne 0 \in \mathbb{R}^{2A}$ such that

$$(L_1 - L_k)^\top = z_{1;k}^\top \begin{bmatrix} S_1 \\ S_k \end{bmatrix}. \tag{29}$$

Now, we have

$$\begin{aligned} & z_{1;k}^\top \begin{bmatrix} S_1 \\ S_k \end{bmatrix} (p - p') \\ & \le \|z_{1;k}\| \left\| \begin{bmatrix} S_1 \\ S_k \end{bmatrix} (p - p') \right\| \quad \text{(by Cauchy-Schwarz ineq.)} \\ & \le \sqrt{2A}\, \|z_{1;k}\| \quad \text{(by Eq. (28))}, \end{aligned} \tag{30}$$

and

$$\begin{aligned} & z_{1;k}^\top \begin{bmatrix} S_1 \\ S_k \end{bmatrix} (p - p') \\ & = (L_1 - L_k)^\top (p - p') \quad \text{(by Eq. (29))} \\ & = (L_1 - L_k)^\top p + (L_k - L_1)^\top p' \\ & \ge \gamma_k \quad \text{(by Condition 1 \& def. of } \gamma_k). \end{aligned} \tag{31}$$

Therefore, from (30) and (31), we have

$$\gamma_k \le \sqrt{2A}\, \|z_{1;k}\|. \tag{32}$$

This inequality does not hold for all $k \ne 1$ for the predefined value of ϵ, since we have

$$\epsilon < \frac{1}{2\sqrt{A}} \min_{k \ne 1} \frac{\gamma_k}{\|z_{1;k}\|}. \tag{33}$$

Therefore, the proof is completed by contradiction. □

Remark. The similar result holds when the optimal action 1 is replaced with action $j \ne k$ such that $\gamma_{j;k} := (L_j - L_k)^\top p > 0$ by taking ϵ satisfying

$$\epsilon < \frac{1}{2\sqrt{A}} \min_{j \ne k:\, \gamma_{j;k} > 0} \frac{\gamma_{j;k}}{\|z_{j;k}\|}. \tag{34}$$

From Lemma 7, we have the following corollary.

Corollary 8. For any $p \in \mathbb{R}^M$ satisfying $p \in C_i$ and $\|S_1 p - S_1 p'\| < \epsilon$, we have

$$\|S_i p - S_i p'\| > \epsilon. \tag{35}$$

Proof. Note that $p \in C_i$ is equivalent to $(L_1 - L_i)^\top p > 0$ for any $i \ne 1$. Therefore, the result directly follows from Lemma 7. □

The next lemma is the property of Mahalanobis distance corresponding to $g_t(p)$.

Lemma 9. Define $T_i := \{p \in R^M : \|S_i p - S_i \hat{p}\| > \epsilon\}$. Assume that $N_i(t) \geq n_i$, $\|S_i \hat{p}_t - S_i p\| \leq \epsilon/4$. Then, for any $0 < \alpha < 1/2$

$$\Pr\left[\inf_{p \in T_i} \|B_t^{1/2}(p - \hat{p}_t)\|^2 \leq \left(\frac{9}{16}\right)^2 n_i (1 - 2\alpha)^{M/2}\right]. \tag{36}$$

Proof. To bound the LHS of the above inequality, we bound $\|B_t^{1/2}(p - \hat{p}_t)\|^2$ from below for $p \in T_i$. Using the triangle inequality and the assumptions, we have

$$\begin{aligned} \|S_i(p - \hat{p}_t)\| &\geq \|S_i p - S_i \hat{p}\| - \|S_i \hat{p}_t - S_i p\| \\ &> \epsilon - \epsilon/4 > 0. \end{aligned} \tag{37}$$

Therefore, we have

$$\begin{aligned} \|B_t^{1/2}(p - \hat{p}_t)\|^2 &\geq \sum_{k \in [N]} N_k(t) \|S_k(p - \hat{p}_t)\|^2 \quad \text{(by def. of } B_t) \\ &\geq n_i \|S_i(p - \hat{p}_t)\|^2 \quad (N_i(t) \geq n_i) \\ &> \left(\frac{9}{16}\right)^2 n_i \quad \text{(by Eq. (37))}. \end{aligned} \tag{38}$$

By the Chernoff bound for a chi-squared random variable in Lemma 6, we now have

$$h(a) \leq e^{-a}(1 - 2\alpha)^{M/2}; \tag{39}$$

for any $a \geq 0$ and $0 < \alpha < 1/2$. Hence, using the fact that $\|B_t^{1/2}(p - \hat{p}_t)\|^2$ follows the chi-squared distribution with M degree of freedom, we have

$$\begin{aligned} \Pr\left[\inf_{p \in T_i} \|B_t^{1/2}(p - \hat{p}_t)\|^2 \leq h\left(\frac{9}{16}\right)^2 n_i\right] \\ \leq \exp\left(-\left(\frac{9}{16}\right)^2 n_i\right)(1 - 2\alpha)^{M/2}; \end{aligned} \tag{40}$$

which completes the proof. □

E.3 Statistics of Uninterested Actions

For any $k \neq i$ and $n_k \in [T]$, define

$$Z_{n_k} := n_k \|q_{k;n_k} - S_k p\|^2; \tag{41}$$

$$Z_{ni} := \sum_{k \neq i} \max_{n_k \in [T]} Z_{n_k}. \tag{42}$$

In this section, we bound $\mathbb{E}[Z_{ni}]$ from above. Note that Z_{ni} is independent of the randomness of Thompson sampling.

Lemma 10 (Upper bound for the expectation of Z_{ni}).

$$\mathbb{E}[Z_{ni}] \leq 4N\left(\log T + \frac{A}{2}\log 2 + 1\right). \tag{43}$$

Proof. Recall that in linear partial monitoring, the feedback $y(t) \in R^A$ for action k is given as

$$y_t = S_k p + \xi, \quad \xi \sim N(0; I_A) \tag{44}$$

at round $t \in [T]$, Therefore $y(t) \sim S_k p + N(0; I_A)$. Since $q_{k;n_k} = \frac{1}{n_k}\sum_{s \in [T]: i(s) = k} y(s)$ for any $n_k \in [T]$, we have

$$q_{k;n_k} - S_k p = \frac{1}{n_k}\sum_{s \in [T]: i(s) = k} (y(s) - S_k p) \sim N(0; I_A/n_k). \tag{45}$$

Therefore,

$$\sqrt{n_k}(\hat{q}_{k;n_k} - S_k p) \Rightarrow N(0; I_A); \tag{46}$$

and thus

$$n_k \|\hat{q}_{k;n_k} - S_k p\|^2 = \|\sqrt{n_k}(\hat{q}_{k;n_k} - S_k p)\|^2 \Rightarrow \chi^2_A. \tag{47}$$

Therefore, for any $0 < 1=2$,

$$\begin{aligned}
E\left[\max_{n_k 2[T]} Z_{n_k}\right] &= \int_0^1 P\left[\max_{n_k 2[T]} Z_{n_k} \quad x\right] dx \\
&\int_0^1 [1 \wedge T \ P\{Z_1 \quad x\}] dx \quad \text{(by the union bound)} \\
&\int_0^1 \left[1 \wedge T \ e^{-x(1-2)^{\frac{A}{2}}}\right] dx \quad \text{by } Z_1 \quad \chi^2_A \text{ and Lemma 6} \\
&= \int_0^x dx + \int_x^1 T \ e^{-x(1-2)^{\frac{A}{2}}} dx \\
&= x + T \int_x^1 e^{-x(1-2)^{\frac{A}{2}}} dx \\
&= x + T(1-2)^{-\frac{A}{2}} \frac{1}{x} e^{-x} \Big|_x^1 \\
&= \frac{1}{n} \log T - \frac{A}{2} \log(1-2) + 1 \Big|_0^x;
\end{aligned} \tag{48}$$

where $x := \frac{1}{n} \log T - \frac{A}{2} \log(1-2)$. Therefore, taking $= 1=4$, we have

$$\begin{aligned}
E\left[\|Z_{ni}\|\right] &= E\left[\sum_{k=1}^{N} \max_{n_k 2[T]} \|Z_{n_k}\|\right] \\
&\sum_{k=1}^{N} E\left[\max_{n_k 2[T]} \|Z_{n_k}\|\right] \\
&(N-1)\frac{1}{n} \log T - \frac{A}{2} \log(1-2) + 1 \\
&4N \log T + \frac{A}{2} \log 2 + 1;
\end{aligned} \tag{49}$$

which completes the proof. □

E.4 Mahalanobis Distance Process

Discussions in this section are essentially very similar to Abbasi-Yadkori et al. (2011, Lemma 11), but their results are not directly applicable and we give the full derivation for self-containedness. To maximize the applicability here we only assume sub-Gaussian noise rather than a Gaussian one.

Let $_t$ be zero-mean 1-sub-Gaussian random variable, which satisfies

$$E\left[e^{\ t}\right] \quad e^{\frac{k\ k^2}{2}} \tag{50}$$

for any $2 R^M$.

Lemma 11. For any vector $v 2 R^M$ and positive definite matrix $V 2 R^{M-M}$ such that $V \quad I$,

$$E_t\left[e^{\frac{k\ t+v k^2_{V^{-1}}}{2}}\right] \quad \frac{p\overline{jVj}}{\sqrt{jV \quad Ij}} e^{\frac{1}{2}v^\top (V-I)^{-1}v}. \tag{51}$$

Proof. For any $x 2 R^M$

$$E_{N(0;V^{-1})}\left[e^{\ x}\right] = e^{\frac{k x k^2_{V^{-1}}}{2}}: \tag{52}$$

Therefore, by letting $x = \ _t + v$ we see that

$$e^{\frac{k \ _t + v k_V^2 \ _1}{2}} = E_{\ N(0;V \ ^1)} \left[e^{\ > (\ _t + v)} \right] : \tag{53}$$

As a result, by the de nition of sub-Gaussian random variables, we have

$$\begin{aligned}
E_t \ e^{\frac{k \ _t + v k_V^2 \ _1}{2}} &= E_{\ N(0;V \ ^1)} \left[E_t \left[e^{\ > (\ _t + v)} \right] \right] \\
&= E_{\ N(0;V \ ^1)} \left[e^{\ > v} E_t \left[e^{\ > \ _t} \right] \right] \\
&\quad E_{\ N(0;V \ ^1)} \left[e^{\ > v} e^{k \ k^2=2} \right] \\
&= \frac{1}{(2 \ )^{d=2} \sqrt{jV \ ^1 j}} \int \ e^{\ > v} e^{k \ k^2=2} e^{k \ k_V^2=2} d \\
&= \frac{1}{(2 \ )^{d=2} \sqrt{jV \ ^1 j}} \int \ e^{\frac{1}{2}(\ > (V \ I) \ \ 2v^> \ )} d \\
&= \frac{\sqrt{jV \ \ I j}}{(2 \ )^{d=2} \sqrt{jV \ ^1 j j V \ \ I j}} \int \ e^{\frac{1}{2}((\ \ (V \ I)^{\ 1} v)^> (V \ I)(\ \ (V \ I)^{\ 1} v) \ \ v^> (V \ I)^{\ 1} v)} d \\
&= \sqrt{\frac{jVj}{jV \ \ I j}} e^{\frac{1}{2} v^> (V \ I)^{\ 1} v} :
\end{aligned} \tag{54}$$

□

Lemma 12.

$$E \ \exp \ \frac{1}{2} \ k\hat{p}_t \ \ p \ k^2_{B_t} \ k \hat{p}_{t \ 1} \ \ p \ k^2_{B_{t \ 1}} \qquad \hat{p}_{t \ 1}; B_{t \ 1}; S_{i(t \ 1)} \qquad \sqrt{\frac{jB_t j}{jB_{t \ 1} j}} : \tag{55}$$

Proof. Let $Z_t := \ \ p \ + \sum_{s=1}^{t} S^>_{i(s)} \ _s$, and we have

$$\begin{aligned}
B_t &= I \ + \sum_{s=1}^{t} S^>_{i(s)} S_{i(s)}, \\
b_t &= \sum_{s=1}^{t} S^>_{i(s)} y(s) = B_t p \ + Z_t, \\
\hat{p}_t &= B_t^{\ 1} b_t = p \ + B_t^{\ 1} Z_t .
\end{aligned}$$

In the following we omit the conditioning on $(\hat{p}_{t \ 1}; B_{t \ 1}; S_{i(t \ 1)})$ for notational simplicity.

Let us de ne $C_t := S_{i(t)} B_{t \ 1} S^>_{i(t)}$ and $d_t := S_{i(t)} B_{t \ 1}^{\ 1} Z_{t \ 1} = S_{i(t)} (\hat{p}_t \ \ p \ )$. Then, using the Sherman-Morrison-Woodbury formula we have

$$
\begin{aligned}
&k\hat{p}_t - p\|_{B_t}^2 - k\hat{p}_{t-1} - p\|_{B_{t-1}}^2 \\
&= Z_t^\top B_t^{-1} Z_t - Z_{t-1}^\top B_{t-1}^{-1} Z_{t-1} \\
&= (Z_{t-1}^\top + \Delta_t^\top S_{i(t)})(B_{t-1}^{-1} - B_{t-1}^{-1} S_{i(t)}^\top (I + S_{i(t)} B_{t-1}^{-1} S_{i(t)}^\top)^{-1} S_{i(t)} B_{t-1}^{-1})(Z_{t-1} + S_{i(t)}^\top \Delta_t) - Z_{t-1}^\top B_{t-1}^{-1} Z_{t-1} \\
&= (Z_{t-1}^\top + \Delta_t^\top S_{i(t)}) B_{t-1}^{-1} (Z_{t-1} + S_{i(t)}^\top \Delta_t) - Z_{t-1}^\top B_{t-1}^{-1} Z_{t-1} \\
&\quad - (Z_{t-1}^\top + \Delta_t^\top S_{i(t)}) B_{t-1}^{-1} S_{i(t)}^\top (I + S_{i(t)} B_{t-1}^{-1} S_{i(t)}^\top)^{-1} S_{i(t)} B_{t-1}^{-1} (Z_{t-1} + S_{i(t)}^\top \Delta_t) \\
&= \Delta_t^\top S_{i(t)} B_{t-1}^{-1} S_{i(t)}^\top \Delta_t + 2 Z_{t-1}^\top B_{t-1}^{-1} S_{i(t)}^\top \Delta_t \\
&\quad - (Z_{t-1}^\top + \Delta_t^\top S_{i(t)}) B_{t-1}^{-1} S_{i(t)}^\top (I + S_{i(t)} B_{t-1}^{-1} S_{i(t)}^\top)^{-1} S_{i(t)} B_{t-1}^{-1} (Z_{t-1} + S_{i(t)}^\top \Delta_t) \\
&= \Delta_t^\top C_t \Delta_t + 2 d_t^\top \Delta_t - (d_t + \Delta_t^\top C_t)(I + C_t)^{-1} (d_t + C_t \Delta_t) \\
&= \Delta_t^\top C_t (I - (I + C_t)^{-1} C_t) \Delta_t + 2 d_t^\top (I - (I + C_t)^{-1} C_t) \Delta_t - d_t^\top (I + C_t)^{-1} d_t \\
&= \Delta_t^\top C_t (I + C_t)^{-1} \Delta_t + 2 d_t^\top (I + C_t)^{-1} \Delta_t - d_t^\top (I + C_t)^{-1} d_t \\
&= \|\Delta_t + C_t^{-1} d_t\|^2_{C_t (I + C_t)^{-1}} - d_t^\top (I + C_t)^{-1} C_t^{-1} d_t - d_t^\top (I + C_t)^{-1} d_t \\
&= \|\Delta_t + C_t^{-1} d_t\|^2_{C_t (I + C_t)^{-1}} - d_t^\top (I + C_t)^{-1} (I + C_t^{-1}) d_t : \qquad (56)
\end{aligned}
$$

Therefore, Lemma 11 with $W := C_t(I + C_t)^{-1} \succeq 0 = (I + C_t)C_t^{-1}$; $v := C_t^{-1} d_t$ yields

$$
\begin{aligned}
&\mathbb{E}\exp\Big(\frac{1}{2}\|k\hat{p}_t - p\|_{B_t}^2 - k\hat{p}_{t-1} - p\|_{B_{t-1}}^2\Big) \\
&\quad \frac{\sqrt{|(I + C_t)C_t^{-1}|}}{\sqrt{|(I + C_t)C_t^{-1} - I|}} e^{\frac{1}{2} d_t^\top C_t^{-1}((I + C_t)C_t^{-1} - I)^{-1} C_t^{-1} d_t} e^{\frac{1}{2} d_t^\top (I + C_t)^{-1}(I + C_t^{-1}) d_t} \\
&\quad \frac{\sqrt{|(I + C_t)C_t^{-1}|}}{\sqrt{|C_t^{-1}|}} e^{\frac{1}{2} d_t^\top C_t^{-1}(C_t^{-1})^{-1} C_t^{-1} d_t} e^{\frac{1}{2} d_t^\top (I + C_t)^{-1}(I + C_t^{-1}) d_t} \\
&= \sqrt{|(I + C_t)|} \\
&= \sqrt{\frac{|B_t|}{|B_{t-1}|}} ; \qquad (57)
\end{aligned}
$$

where see, e.g., Abbasi-Yadkori et al. (2011, Lemma 11) for the last equality. □

E.5 Norms under Perturbations

In the following two lemmas, we give some analysis of norms under perturbations.

Lemma 13. Let A be a positive definite matrix. Let $a \in \mathbb{R}^d$ and $\gamma > 0$ be such that $\gamma < \|a\| = 3\gamma$. Then

$$
\min_{x : \|x\| \le \gamma} \max_{x' : \|x'\| \le \gamma} (a + x + x')^\top A (a + x + x') = \min_{x'' : \|x''\| \le \gamma} (a + x'')^\top A (a + x'') : \qquad (58)
$$

Proof. By considering the Lagrangian multiplier we see that any stationary point of the function $(a + x'')^\top A(a + x'')$ over $\{(x, x') : \|x\| \le \gamma,\ \|x'\| \le \gamma\}$ satisfies

$$
\begin{aligned}
&A(a + x + x') - \lambda_1 x = 0 ; \\
&A(a + x + x') - \lambda_2 x' = 0 ; \\
&x^\top x = 4\gamma^2 ; \\
&x'^\top x' = \gamma^2 ; \qquad (59)
\end{aligned}
$$

and therefore $\lambda_1 x = \lambda_2 x'$. Considering the last two conditions of (59) we have $\lambda_2 = 2\lambda_1$, implying that

$$
x' = -(3A - 2\lambda_1 I) A a \qquad (60)
$$

or

$$x^0 = (A - \lambda_1 I)^{\dagger} A a \tag{61}$$

for λ_1 satisfying $x^{0\top} x^0 = r^2$.

Note that it holds for any positive definite matrix B that

$$\frac{d^2}{d\lambda^2} a^\top (B + \lambda I)^{-2} a = a^\top (B + \lambda I)^{-4} a = \| (B + \lambda I)^{-2} a \|^2, \tag{62}$$

which is positive almost everywhere, meaning that $a^\top (B + \lambda I)^{-2} a$ is strictly convex with respect to $\lambda \in \mathbb{R}$. Therefore, there exists at most two λ's satisfying (60) and $x^{0\top} x^0 = r^2$, and there exists at most two λ_1's satisfying (61) and $x^{0\top} x^0 = r^2$. In summary, there at most four stationary points of $(a + x'')^\top A (a + x'')$ over $\{(x, x^0) : \|x\| \le \alpha_2, \|x^0\| \le \alpha_3\}$.

On the other hand, two optimization problems

$$\min_{x : \|x\| \le \alpha_2} \min_{x^0 : \|x^0\| \le \alpha_3} (a + x + x^0)^\top A (a + x + x^0) = \min_{x'' : \|x''\| \le \alpha_3} (a + x'')^\top A (a + x'') \tag{63}$$

and

$$\max_{x : \|x\| \le \alpha_2} \max_{x^0 : \|x^0\| \le \alpha_3} (a + x + x^0)^\top A (a + x + x^0) = \max_{x'' : \|x''\| \le \alpha_3} (a + x'')^\top A (a + x'') \tag{64}$$

can be easily solved by an elementary calculation and the optimal values are equal to those corresponding to (60).

Therefore, the optimal solutions of the two minimax problems

$$\max_{x : \|x\| \le \alpha_2} \min_{x^0 : \|x^0\| \le \alpha_3} (a + x + x^0)^\top A (a + x + x^0) \tag{65}$$

and

$$\min_{x : \|x\| \le \alpha_2} \max_{x^0 : \|x^0\| \le \alpha_3} (a + x + x^0)^\top A (a + x + x^0) \tag{66}$$

correspond to two points corresponding to (61).

We can see again from an elementary calculation that the optimal solutions for two optimization problems

$$\begin{aligned} &\min_{x'' : \|x''\| \le \alpha_3} (a + x'')^\top A (a + x'') \\ &\max_{x'' : \|x''\| \le \alpha_3} (a + x'')^\top A (a + x'') \end{aligned} \tag{67}$$

have the same necessary and sufficient conditions (61) and we complete the proof by noticing that (65) is less than (66). □

Lemma 14. Let $A = nS_1^\top S_1$ be a positive-definite matrix with minimum eigenvalue at least ε. Then, for any $\hat{p} \in \mathbb{R}^d$ and $\alpha > 0$ satisfying $\alpha < \|\hat{p} - p\|$,

$$\|\hat{p} - p\|_A^2 \inf_{p : \|p - \hat{p}\| = \alpha} \sup_{p^0 : \|p^0 - p\| \le \alpha} \|p^0 - \hat{p}\|_A^2 \ge \sqrt{n}\, \varepsilon \alpha \|\hat{p} - p\|. \tag{68}$$

Proof. Let $a = \hat{p} - p$. By Lemma 13, we have

$$\begin{aligned} &\inf_{p : \|p - \hat{p}\| = \alpha} \sup_{p^0 : \|p^0 - p\| \le \alpha} \|p^0 - \hat{p}\|_A^2 \\ &= \inf_{x : \|x\| = \alpha} \sup_{x^0 : \|x^0\| \le \alpha} \|a + x + x^0\|_A^2 \\ &= \inf_{x : \|x\| = \alpha} \|a + x\|_A^2, \end{aligned} \tag{69}$$

Now define $S_{0;A} = \{x : \|x\|_A \le 0\}$. Then, we see that $S_{\bar{p};A} = \{x : \|x\| \le 0\}$. Therefore, an elementary calculation using the Lagrange multiplier technique shows

$$\inf_{x:\|x\|\le} \|p' - \hat{p}\|_A^2 \le \inf_{x \in S_{\bar{p};A}} \|p - \hat{p}\|_A^2 = \|a\|_A^{\bar{p}^2}. \tag{70}$$

As a result, we see that

$$\begin{aligned}
\|p - \hat{p}\|_A^2 \le \inf_{p:\|p-p\|\le 2} \sup_{p':\|p'-p\|\le} \|p' - \hat{p}\|_A^2 &\le \|a\|_A^2 \le \|a\|_A^{\bar{p}^2} \\
&= \Big(\sqrt{\bar{p}}\|a\|_A + \|a\|_A\Big)\sqrt{\bar{p}} \\
&= \Big(\sqrt{\bar{p}}\|a\|_A + \|a\|\Big)\sqrt{\bar{p}} \\
&= \sqrt{\bar{p}}\|a\|_A + \sqrt{\bar{p}}(\|a\| \quad) \\
&\le \sqrt{\bar{p}}\|a\|_A \\
&\le \sqrt{n}\,\|S_1 a\|.
\end{aligned} \tag{71}$$

□

For the subsets of $\mathbb{R}^n$, X and Y, let $X + Y := \{x + y : x \in X, y \in Y\}$ be the Minkowski sum, and let $B_r^n(p)$ be the n-dimensional Euclidian ball of radius r at point $p \in \mathbb{R}^n$ (the superscript n can be omitted when it is clear from context). We also let δ_0 be

$$\delta_0 := \frac{}{16 \max_{i\in[N]} \|S_i\| \quad \sqrt{\frac{1}{A}} \max_{i\in[N]} \frac{\|L_i - L_1\|}{\|z_{1;i}\|}}; \tag{72}$$

which is also used throughout the proof of this appendix and Appendix F as in (26).

Theorem 15. Let $\delta'' \in (0,)$ be a constant defined in (26). Let $\hat{p} \in C_k + B_{\delta_0}^d(0)$ be satisfying $\|S_k(\hat{p} - p)\| \le \delta''$. Then, there exists > 0 satisfying for any $n \ge 0$ and $A \ge nS_1^{\ge} S_1 + I$ that

$$\|p - \hat{p}\|_A^2 \le \inf_{p:\|p-p\|\le 2} \sup_{p':\|p'-p\|\le} \|p' - \hat{p}\|_A^2 \le \sqrt{n}. \tag{73}$$

Proof. Recall that $\delta'' < \min_{p \in C_1^c} \|p - p\| = 3$. It is enough from Lemma 14 to prove that

$$:= \min_{\hat{p}\in\{p\in C_k + B_{\delta_0}^d(0):\ \|S_k(p - p)\| \le \delta''\}} \|S_1(\hat{p} - p)\| \tag{74}$$

is positive.

We prove by contradiction and the proof is basically same as that of Lemma 7 but more general in the sense that the condition $\hat{p}$ is not $\hat{p} \in C_k$ but $\hat{p} \in C_k + B_{\delta_0}^d(0)$. Assume that $= 0$, that is, there exists $\hat{p} \in C_k + B_{\delta_0}^d(0)$ satisfying $\|S_k(p - p)\| \le \delta''$ and $\|S_1(\hat{p} - p)\| = 0$. Note that $\|S_1(\hat{p} - p)\| = 0$ implies $\|S_1(\hat{p} - p)\| \le \delta''$. Therefore, we now have following conditions on $\hat{p}$:

$$\begin{aligned}
&\hat{p} \in C_k + B_{\delta_0}^d(0) \\
&\|S_1(\hat{p} - p)\| \le \delta'' \\
&\|S_k(\hat{p} - p)\| \le \delta''.
\end{aligned}$$

Following the same argument as the proof of Lemma 7, we have

$$z_{1;k}^{>} \frac{S_1}{S_k}(\hat{p} - p) \le \sqrt{2A}\,\delta''\|z_{1;k}\|. \tag{75}$$

On the other hand, since $\hat{p} \in C_k + B^{d}_{0}(0)$ we can take $p \in C_k$ such that $\|\hat{p} - p\| < \delta'$. Hence,

$$z_{1;k}^{\top} \begin{matrix} S_1 \\ S_k \end{matrix} (\hat{p} - p) = (L_1 - L_k)^{\top}(\hat{p} - p)$$

$$= (L_k - L_1)^{\top}(\hat{p} - p) + (L_1 - L_k)^{\top} p + (L_k - L_1)^{\top} p$$

$$(L_k - L_1)^{\top}(\hat{p} - p) + \varepsilon_k : \quad (\text{by } p \in C_k \text{ and def. of } \varepsilon_k) \tag{76}$$

From (75) and (76), we have

$$\varepsilon_k \quad (L_k - L_1)^{\top}(\hat{p} - p) \quad \sqrt{2A''}\|z_{1;k}\| : \tag{77}$$

Now, the left hand side of (77) is bounded from below as

$$\varepsilon_k \quad (L_k - L_1)^{\top}(\hat{p} - p) \quad \varepsilon_k \| L_k - L_1\| \|\hat{p} - p\|$$

$$\varepsilon_k \| L_k - L_1\| \, \delta'$$

$$= \varepsilon_k \| L_k - L_1\| \frac{\sqrt{\frac{1}{A}} \max_i \frac{\|L_1 - L_i\|}{\|z_{1;i}\|}}$$

$$= \varepsilon_k \| L_k - L_1\| \frac{\frac{1}{2}\sqrt{\frac{1}{A}} \min_i \frac{i}{\|z_{1;i}\|}}{\sqrt{\frac{1}{A}} \max_i \frac{\|L_1 - L_i\|}{\|z_{1;i}\|}}$$

$$\varepsilon_k \quad \varepsilon_k = 2 : \tag{78}$$

On the other hand, using the de nition of δ'', the right hand side of (77) is bounded from above as

$$\sqrt{2A''}\|z_{1;k}\| < \varepsilon_k = 2 : \tag{79}$$

Therefore, the proof is completed by contradiction. □

E.6 Exit Time Analysis

We next consider the exit time. Let A_t be an event deterministic given F_t, and B_t be a random event such that if B_t occurred then $A_{t'}$ never occurs for $t' = t+1, t+2, \ldots$. Let $P_t, t = 1, 2, \ldots, T$, be a stochastic process satisfying $P_t = \Pr\{B_t | F_t\}$ a.s. and P_t^{-1} is a supermartingale with respect to the filtration induced by F_t.

Theorem 16. Let τ be the stopping time defined as

$$\tau = \begin{cases} \min\{t \in [T] : A_t\} & \text{if } A_t \text{ occurs for some } t \in [T]. \\ T+1 & \text{otherwise} \end{cases} \tag{80}$$

Then we almost surely have

$$E\left[\sum_{t=1}^{T} 1[A_t] \;\middle|\; F_0\right] \quad P_0^{-1} \quad T; \quad \tau = T+1 : \tag{81}$$

We prove this theorem based on the following lemma.

Lemma 17. Let $(Q_i)_{i=1}^{\infty}$ $[0,1]$ be an arbitrary stochastic process such that $(Q_t^{-1})_{i=1}^{\infty}$ is a supermartingale with respect to a filtration $(G_i)_{i=1}^{\infty}$. Then, for any $G_0 \quad G_1$,

$$E\left[\sum_{i=1}^{T}\prod_{j=1}^{i}(1 - Q_j) \;\middle|\; G_0\right] \quad E\left[Q_1^{-1} | G_0\right] \quad 1 \quad \text{a.s.} \tag{82}$$

Proof. Let

$$N_k((Q_i, G_i)_{i=1}^{\infty}, G_0) = E\left[\sum_{i=1}^{k}\prod_{j=1}^{i}(1 - Q_j) \;\middle|\; G_0\right]$$

$$\overline{N}_k((Q_i, G_i)_{i=1}^{\infty}, G_0) = E\left[\sum_{i=1}^{t}\prod_{j=1}^{i}(1 - Q_j) \;\middle|\; G_0\right] \quad \text{where} Q_j = Q_k \text{ for } j > k . \tag{83}$$

We show $\overline{N}_k((Q_i;G_i)_{i=1}^{1};G_0) \quad E[Q_1^{\ 1}|G_0] \quad 1$ a.s. for any $(Q_i;G_i)_{i=1}^{1}$, $G_0 \quad G_1$ and $k \ 2 \ N$ by induction. First, for $k = 1$ the statement holds since

$$\begin{aligned}\overline{N}_1((Q_i;G_i)_{i=1}^{1};G_0) &= E\Big[\sum_{i=1}^{1}\prod_{j=1}^{i}(1 \quad Q_1) \quad G_0\Big] \\ &= E\left[Q_1^{\ 1} \quad 1 \quad G_0\right] \\ &= E\left[Q_1^{\ 1} \quad G_0 \quad 1\right] \end{aligned} \tag{84}$$

Next, assume that the statement holds for all $(Q_i;G_i)_{i=1}^{k}$, $G_0 \quad G_1$ and $k \quad k_0$. Then, we almost surely have

$$\begin{aligned}\overline{N}_{k_0+1}((Q_i;G_i)_{i=1}^{1};G_0) &= E\Big[(1 \quad Q_1)E\Big[1+\sum_{i=2}^{1}\prod_{j=2}^{i}(1 \quad Q_j) \quad G_1\Big] \quad G_0\Big] \\ &= E\left[(1 \quad Q_1)(1+\overline{N}_{k_0}((Q_i;G_i)_{i=2}^{1};G_1)) \quad G_0\right] \\ &\quad E\left[(1 \quad Q_1)E[Q_2^{\ 1} \quad G_1] \quad G_0\right] \quad \text{(assumption of the induction)} \\ &\quad E\left[Q_1^{\ 1} \quad G_0\right] \quad 1 \quad Q_i^{\ 1} \text{ is a supermartingale.}\end{aligned} \tag{85}$$

We obtain the lemma from

$$E\Big[\sum_{i=1}^{k}\prod_{j=1}^{i}(1 \quad Q_j) \quad G_0\Big] = N_k((Q_i;G_i)_{i=1}^{1};G_0) \quad \overline{N}_k((Q_i;G_i)_{i=1}^{1};G_0) \quad \text{a.s.} \tag{86}$$

□

Proof of Theorem 16. The statement is obvious for the case $= T + 1$ and we consider the other case in the following.

Let $_i$ be the time of the i-th occurrence of A_t. More formally, we de ne $_i$ as the stopping time $_1 =$ and

$$_{i+1} = \begin{cases} \min\left\{t \ 2 \ [T] : \sum_{t'=1}^{T} 1[A_{t'}] = i+1\right\} & \sum_{t'=1}^{T} 1[A_{t'}] \quad i+1; \\ _i + 1 & \text{otherwise.}\end{cases} \tag{87}$$

Then $(P_i^0) = (P_{\ i})$ is a stochastic process measurable by the ltration induced by $(F_i^0) = (F_{\ i})$. By Lemma 17 we obtain

$$\begin{aligned}E\Big[\sum_{t=1}^{T}1[A_t] \quad F\Big] &= E\Big[\sum_{n=1}^{T}1\sum_{t=1}^{T}1[A_t] \quad n \quad F\Big] \\ &\quad 1+E\Big[\sum_{n=2}^{T}1\sum_{t=1}^{T}1[A_t] \quad n \quad F\Big] \\ &\quad 1+E\Big[\sum_{i=1}^{T}\prod_{j=1}^{i}(1 \quad P_j^0) \quad F_1^0\Big] \\ &\quad 1+E\left[(P_1^0)^{\ 1}|F_1^0 \quad 1\right] \\ &= P^{\ 1}:\end{aligned} \tag{88}$$

□

F Regret Analysis of TSPM Algorithm

In this appendix, we give the proof of Theorem 3. Note that the cells are de ned for the decomposition of R^M, not P_M. In other words, the cell C_i is here de ned as $C_i = \{p \ 2 \ R^M : \text{action } i \text{ is optimal}\}$.

For the linear setting, the empirical feedback distribution $q_i^{(t)}$ and $q_{i,n}$ are defined as

$$q_i^{(t)} := \frac{1}{N_i(t)} \sum_{s \in [t-1]: i(s)=i} y(s), \tag{89}$$

$$q_{i,n} := \text{the value of } q_i^{(t)} \text{ after taking action } i \text{ for } n \text{ times.} \tag{90}$$

Recall that $\hat{p}_t = B_t^{-1} b_t$, which is the mode of $g_t(p)$.

F.1 Regret Decomposition

Here, we break the regret into several terms. For any $i \in [N]$, we define events

$$A_i(t) := \left\{ \|\hat{p}_t - S_i\|_2 \le \frac{\Delta}{4} \right\}, \tag{91}$$

$$\tilde{A}_i(t) := \{ \|\hat{p}_t - S_i\|_2 \le \Delta \}. \tag{92}$$

We first decompose the regret as

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \Delta_{i(t)} \\
&\le \sum_{t=1}^{T} \Delta_{i(t)} \left[1\{\tilde{A}_1(t)\} + \max_{j \in [N]} 1\{\tilde{A}_1^c(t)\} \right] \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T} \Delta_i 1\{i(t) = i, \tilde{A}_1(t)\} + \max_{j \in [N]} \sum_{t=1}^{T} 1\{\tilde{A}_1^c(t)\} \\
&\le \sum_{i=1}^{N} \sum_{t=1}^{T} \Delta_i \underbrace{1\{i(t) = i, \tilde{A}_1(t), A_i(t)\}}_{\text{(A)}} + \sum_{i=1}^{N} \Delta_i \underbrace{1[i(t) = i, A_i^c(t)]}_{\text{(B)}} + \max_{j \in [N]} \sum_{t=1}^{T} 1\{\tilde{A}_1^c(t)\}.
\end{aligned} \tag{93}$$

To decompose the last term, we define the following notation. We define for any $i \in [N]$

$$P_i(t) := P\{\hat{p}_t \in C_i \mid \mathcal{F}_t\}. \tag{94}$$

We also define

$$C_{i,t} := C_i \setminus B_{\epsilon_0}(\hat{p}_t), \tag{95}$$

where ϵ_0 is defined in (72), and

$$i_t := \arg\max_{i \in [N]} P\{\hat{p}_t \in C_{i,t} \mid \mathcal{F}_t\}. \tag{96}$$

We define p_t as an arbitrary point in $C_{i_t,t}$. Then, we define

$$A_i(t) := \left\{ \|p_t - S_i\|_2 \le \frac{\Delta}{8} \right\}. \tag{97}$$

Using these notations, the last term in (93) can be decomposed as

$$\begin{aligned}
&1\{\tilde{A}_1^c(t)\} \sum_{k=1}^{N} 1\{p_t \in C_k, \tilde{A}_1^c(t)\} \\
&= \sum_{k=1}^{N} 1\{p_t \in C_k, A_k^c(t), \tilde{A}_1^c(t)\} + \sum_{k=1}^{N} 1\{p_t \in C_k, A_k(t), \tilde{A}_1^c(t)\} \\
&\le \sum_{k=1}^{N} \underbrace{1\{p_t \in C_k, A_k^c(t)\}}_{\text{(C)}} + \underbrace{1\{p_t \in C_1, A_1(t), \tilde{A}_1^c(t)\}}_{\text{(D)}} + \sum_{k=2}^{N} \underbrace{1\{p_t \in C_k, A_k(t)\}}_{\text{(E)}}.
\end{aligned} \tag{98}$$

We will bound the expectation of each term in the following and complete the proof of Theorem 3 as

$$\begin{aligned}
E[\mathrm{Reg}(T)] &= \sum_{i \neq 1} \Delta_i \left(O\left(\frac{1}{\Delta_i^2}\right) \log T \right) + O\left(\frac{N}{\Delta_2^2}\right) \log T \\
&\quad + \max_{j \in [N]} \Delta_j \left(\sum_{k=1}^{N} O\left(\frac{NM}{\Delta_k^2}\right) \log T + O(1) + \sum_{k=2}^{N} O(1) \right) \\
&= O\left(\max\left(\frac{N}{\Delta_{i \in [N]}^2 \Delta_i}, \frac{N^2 M \max_{i \in [N]} \Delta_i}{\Delta_{i \in [N]}^2} \right) \log T \right) \\
&= O\left(\frac{AN^2 M \max_{i \in [N]} \Delta_i}{\Delta_{i \in [N]}^2} \log T \right); \qquad (99)
\end{aligned}$$

where the last transformation follows from the definition of A in (26).

F.2 Analysis for Case (A)

Lemma 18. For any $i \neq 1$,

$$E\left[\sum_{t=1}^{T} \mathbb{1}\left[i(t) = i, \tilde{A}_1(t), A_i(t) \right] \right] \leq \frac{64}{9\Delta_i^2} \log T + 2^{M+2}. \qquad (100)$$

To prove Lemma 18, we prove the following lemma using Corollary 8 and Lemma 9.

Lemma 19. For any $0 < \delta < 1/2$,

$$P\{\hat{p}_t \in V_i \mid A_i(t), N_i(t) > n_i\} \leq \exp\left(-\frac{9}{16} \Delta_i^2 n_i (1 - \delta)^{M+2} \right); \qquad (101)$$

where $V_i := \{p \in C_i : \|S_1 p - S_1 p_*\| \geq \Delta_i/4\}$.

Proof. Since $\hat{p}_t \sim N(\hat{\mu}_t, B_t^{-1})$ for $\hat{\mu}_t = B_t^{-1} b_t$, the squared Mahalanobis distance $\|B_t^{1/2}(p_t - \hat{\mu}_t)\|^2$ follows the chi-squared distribution with M degree of freedom. Therefore, we have

$$P\{\hat{p}_t \in V_i \mid A_i(t), N_i(t) > n_i\} \leq h\left(\inf_{p \in V_i} \|B_t^{1/2}(p - \hat{\mu}_t)\|^2 \right); \qquad (102)$$

where $h(a) = P_{X \sim \chi^2_M}\{X \geq a\}$. To use Lemma 9, we check the condition of Lemma 9 is indeed satisfied. First, it is obvious that the assumptions $N_i(t) > n_i$ and $\|S_i \hat{\mu}_t - S_i p_*\| < \Delta_i/4$ are satisfied. Besides, $p \in V_i$ implies $p \in T_i = \{p \in R^M : \|S_i p - S_i p_*\| \geq \Delta_i/4\}$ from Corollary 8. Thus, applying Lemma 9 concludes the proof. □

Proof of Lemma 18. For any $n_i > 0$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{1}\left[i(t) = i, \tilde{A}_1(t), A_i(t) \right] \\
&= \sum_{t=1}^{T} \mathbb{1}\left[i(t) = i, \tilde{A}_1(t), A_i(t), N_i(t) \leq n_i \right] + \sum_{t=1}^{T} \mathbb{1}\left[i(t) = i, \tilde{A}_1(t), A_i(t), N_i(t) > n_i \right] \\
&\leq n_i + \sum_{t=1}^{T} \mathbb{1}\left[i(t) = i, \tilde{A}_1(t), A_i(t), N_i(t) > n_i \right]. \qquad (103)
\end{aligned}$$

The second term is bounded from above as

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T} \mathbb{1}\left[i(t)=i,\ \tilde{A}_1(t),\ A_i(t),\ N_i(t)>n_i\right]\right] \\
&=\sum_{t=1}^{T} P\left\{i(t)=i,\ \tilde{A}_1(t),\ A_i(t),\ N_i(t)>n_i\right\} \\
&\le \sum_{t=1}^{T} P\left\{i(t)=i,\ \tilde{A}_1(t)\ \ A_i(t),\ N_i(t)>n_i\right\} \\
&=\sum_{t=1}^{T} P\left\{i(t)=i,\ \tilde{A}_1(t),\ \hat{p}_t \in C_i\ \ A_i(t),\ N_i(t)>n_i\right\} \qquad (i(t)=i \text{ implies } \hat{p}_t \in C_i) \\
&\le \sum_{t=1}^{T} P\{\hat{p}_t \in V_i \mid A_i(t),\ N_i(t)>n_i\}. \qquad (104)
\end{aligned}
$$

To obtain an upper bound for $P\{\hat{p}_t \in V_i \mid A_i(t),\ N_i(t)>n_i\}$, we use Lemma 19. By taking $n_i = \frac{16}{9}\frac{1}{\epsilon^2}\log T$ with $\epsilon = 1/4$, we have

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T} \mathbb{1}\left[i(t)=i,\ \tilde{A}_1(t),\ A_i(t)\right]\right] \le n_i + \sum_{t=1}^{T} P\{\hat{p}_t \in V_i \mid A_i(t),\ N_i(t)>n_i\} \\
&\le n_i + \sum_{t=1}^{T} \exp\left(-\frac{9}{16}\epsilon^2 n_i\right)(1-\epsilon)^{M-2} \qquad \text{(by Lemma 19)} \\
&= \frac{16}{9}\frac{1}{\epsilon^2}\log T + (1-\epsilon)^{M-2} \\
&= \frac{64}{9\epsilon^2}\log T + 2^{M-2}; \qquad (105)
\end{aligned}
$$

which completes the proof. □

F.3 Analysis for Case (B)

Lemma 20. For any $i \neq 1$,

$$
E\left[\sum_{t=1}^{T} \mathbb{1}[i(t)=i,\ A_i^c(t)]\right] \le \frac{256N\left(\log T + \frac{A}{2}\log 2 + 1\right)}{\epsilon^2} + \frac{16A^2}{\epsilon^2} \qquad (106)
$$

The regret in this case can intuitively be bounded because as the round proceeds the event $A_i^c(t)$ makes $S_i\hat{p}_t$ close to $S_i p^*$, which implies that the expected number of times the event $A_i^c(t)$ occurs is not large.

Before going to the analysis of Lemma 20, we prove useful inequalities between $\|q_i^{(t)} - S_i p^*\|$, $\|q_i^{(t)} - S_i\hat{p}_t\|$, and $\|S_i\hat{p}_t - S_i p^*\|$.

Lemma 21. Assume $N_i(t) > 0$. Then,

$$
\|q_i^{(t)} - S_i\hat{p}_t\|^2 \le \frac{Z_{n_i}}{N_i(t)} + \|q_i^{(t)} - S_i p^*\|^2. \qquad (107)
$$

Proof. Recall that $\hat{p}_t$ is the maximizer of $g_t(p)$, and we have

$$
\hat{p}_t = \arg\max_{p \in R^M} g_t(p) = \arg\max_{p \in R^M} \prod_{i=1}^{N} \exp\left\{\frac{1}{2}N_i(t)\|q_i^{(t)} - S_i p\|^2\right\} = \arg\min_{p \in R^M} \sum_{i=1}^{N} N_i(t)\|q_i^{(t)} - S_i p\|^2. \qquad (108)
$$

Using this and the definition of Z_{ni}, we have

$$N_i(t)\|q_i^{(t)} - S_i p_t\|^2 \le \sum_{k \in [N]} N_k(t)\|q_k^{(t)} - S_k p_t\|^2 + \sum_{k \in [N]} N_k(t)\|q_k^{(t)} - S_k p\|^2 + Z_{ni} + N_i(t)\|q_i^{(t)} - S_i p\|^2. \tag{109}$$

Dividing by $N_i(t)$ on the both sides completes the proof. □

Lemma 22. Assume that $A_i^c(t)$ and $N_i(t) > 0$ hold. Then,

$$\|q_i^{(t)} - S_i p\| > \frac{1}{2} - \sqrt{\frac{Z_{ni}}{N_i(t)}}. \tag{110}$$

Proof. By the triangle inequality,

$$\begin{aligned}\|q_i^{(t)} - S_i p\| + \|S_i p - S_i p_t\| &\ge \sqrt{\frac{\|S_i p - q_i^{(t)}\|^2}{\frac{Z_{ni}}{N_i(t)} + \|q_i^{(t)} - S_i p\|^2}} \quad \text{(by } A_i^c(t) \text{ and Lemma 21)} \\ &> \frac{1}{4} - \sqrt{\frac{Z_{ni}}{N_i(t)}}\,\|q_i^{(t)} - S_i p\| \quad \text{by } \frac{\sqrt{x+y}}{\sqrt{x}+\sqrt{y}} \text{ for } x, y \ge 0,\end{aligned} \tag{111}$$

which is equivalent to (110). □

Proof of Lemma 20. We first bound the expectation conditioned on Z_{ni}, and then take the expectation for Z_{ni}. Now,

$$\begin{aligned}&E\left[\sum_{t=1}^{T} 1[i(t) = i,\ A_i^c(t)] \,\middle|\, Z_{ni}\right] \\ &= E\left[\sum_{t=1}^{T} 1\left[i(t) = i,\ A_i^c(t),\ N_i(t) \le \frac{64 Z_{ni}}{\epsilon^2}\right] \,\middle|\, Z_{ni}\right] \\ &\quad + E\left[\sum_{t=1}^{T} 1\left[i(t) = i,\ A_i^c(t),\ N_i(t) > \frac{64 Z_{ni}}{\epsilon^2}\right] \,\middle|\, Z_{ni}\right] \\ &\le \frac{64 Z_{ni}}{\epsilon^2} + E\left[\sum_{t=1}^{T} 1\left[i(t) = i,\ A_i^c(t),\ N_i(t) > \frac{64 Z_{ni}}{\epsilon^2}\right] \,\middle|\, Z_{ni}\right] \quad (i(t) = i \text{ for all } t \in [T]).\end{aligned} \tag{112}$$

The first term becomes $256 N \log T + \frac{A}{2}\log 2 + 1 = \epsilon^2$ by taking expectation over Z_{ni} using Lemma 10. Then, we bound the second term. From Lemma 22, $A_i^c(t)$ and $N_i(t) > \frac{64 Z_{ni}}{\epsilon^2}$ im-

ply $k q_i^{(t)} - S_i p\ k > = 16$. Therefore,

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T} 1\{i(t)=i;\ A_i^c(t);\ N_i(t) > \frac{64 Z_{ni}}{2} - Z_{ni}\right] \\
&E\left[\sum_{t=1}^{T} 1\{i(t)=i;\ k q_i^{(t)} - S_i p\ k > \frac{}{16}\right] \\
&E\left[\sum_{t=1}^{T} 1\{i(t)=i;\ \left[\sum_{y\in[A]} j(q_i^{(t)})_y - (S_i)_y p\ j > \frac{p}{16}\frac{}{A}\right]\right] \\
&E\left[\sum_{y=1}^{A}\sum_{t=1}^{T} 1\{i(t)=i;\ j(q_i^{(t)})_y - (S_i)_y p\ j > \frac{p}{16}\frac{}{A}\right] \\
&E\left[\sum_{y=1}^{A}\sum_{n_i=1}^{T}\sum_{t=1}^{T} 1\{i(t)=i;\ N_i(t)=n_i;\ j(q_i^{(t)})_y - (S_i)_y p\ j > \frac{p}{16}\frac{}{A}\right] \\
&= E\left[\sum_{y=1}^{A}\sum_{n_i=1}^{T} 1\left[\sum_{t=1}^{T} i(t)=i;\ N_i(t)=n_i;\ j(q_i^{(t)})_y - (S_i)_y p\ j > \frac{p}{16}\frac{}{A}\right]\right]
\end{aligned}
$$

(The event $\{i(t)=i;\ N_i(t)=n_i\}$ occurs at most once for t and i.)

$$
\begin{aligned}
&\sum_{y=1}^{A}\sum_{n_i=1}^{T} P\ j(q_{i;n_i})_y - (S_i)_y p\ j > \frac{p}{4}\frac{}{A} \\
&\sum_{y=1}^{A}\sum_{n_i=1}^{T} 2\exp\left(-2n_i\left(\frac{p}{4}\frac{}{A}\right)^2\right) \quad \text{(by Hoeffding's ineq.)} \\
&2A\sum_{n_i=1}^{1}\exp\left(-\frac{n_i\ ^2}{8A}\right) \\
&= 2A\frac{1}{\exp\left(\frac{^2}{8A}\right) - 1} \\
&2A\frac{1}{\frac{^2}{8A}} \quad (\text{by } e^x \ge 1+x) \\
&= \frac{16A^2}{^2}: \qquad (113)
\end{aligned}
$$

By summing up the above argument, the proof is completed. □

F.4 Analysis for Case (C)

Before going to the analysis of cases (C), (D), and (E), we recall some notations. Recall that

$$P_i(t) = P\{p_t \in C_i \mid F_t\}; \qquad (114)$$

$C_{i;t} = C_i \setminus B_0(\hat{p}_t)$, $i_t = \arg\max_{i\in[N]} P\{p_t \in C_{i;t} \mid F_t\}$, and p_t is an arbitrary point in $C_{i_t;t}$. Also recall that

$$A_i(t) = \left\{k S_i p_t - S_i p\ k - \frac{}{8}\right\}: \qquad (115)$$

Lemma 23. For any $i \in [N]$,

$$E\left[\sum_{t=1}^{T} 1\{p_t \in C_i;\ A_i^c(t)\right] - \frac{N}{p_0}\left(\frac{2^5 M \log T}{^2} + e^{-kp\ k^2/2}\right)\left(\frac{1}{T} + \frac{L}{M}\right)^{M/2}\left(\frac{1}{1 - e^{-^2/2^5}}\right)^{!}: \qquad (116)$$

Before proving the above lemma, we give two lemmas.

Lemma 24.

$$P\{\hat{p}_t \in C_{i_t} \mid F_t\} \le P\{\hat{p}_t \in C_{i_t;t} \mid F_t\} \le p_0/N; \tag{117}$$

where $p_0 := 1 - h((\epsilon_0)^2)$.

Proof. First, we prove

$$P\left\{\hat{p}_t \in \bigcup_{i\in[N]} C_{i;t} \mid F_t\right\} \le 1 - h((\epsilon_0)^2): \tag{118}$$

This follows from

$$\begin{aligned} P\left\{\hat{p}_t \in \bigcup_{i\in[N]} C_{i;t} \mid F_t\right\} &= P\{\hat{p}_t \in B_{\epsilon_0}(\hat{p}_t) \mid F_t\} \\ &\le \inf_{p\in\{p:\|p-\hat{p}_t\|>\epsilon_0\}} \|B_t^{1/2}(p-\hat{p}_t)\|^2 \\ &\le h\left(\|p-\hat{p}_t\|^2\right) \\ &\le h((\epsilon_0)^2): \end{aligned} \tag{119}$$

Using the de nition of i_t completes the proof. □

Lemma 25. For any $i \in [N]$, the event $A_i^c(t)$ implies $\|S_i\hat{p}_t - S_i p\| \ge \epsilon/16$.

Proof. Using the triangle inequality, we have

$$\begin{aligned} \|S_i\hat{p}_t - S_i p\| &\ge \|S_i p_t - S_i p\| - \|S_i p_t - S_i\hat{p}_t\| \\ &= 8\,\|S_i\|\,\|p_t - \hat{p}_t\| \\ &= 8\,\|S_i\|\frac{\|B_t^{1/2}(p_t-\hat{p}_t)\|}{16\max_i\|S_i\|} \\ &= 8\,\|B_t^{1/2}(p_t-\hat{p}_t)\|/16 = \|B_t^{1/2}(p_t-\hat{p}_t)\|/2: \end{aligned} \tag{120}$$

□

Proof of Lemma 23. For any n_0, which is speci ed later, we have

$$\begin{aligned} &E\left[\sum_{t=1}^T 1\{\hat{p}_t \in C_i, A_i^c(t)\}\right] \\ &= E\left[\sum_{t=1}^T 1\{\hat{p}_t \in C_i, A_i^c(t), N_i(t) < n_0\}\right] + E\left[\sum_{t=1}^T 1\{\hat{p}_t \in C_i, A_i^c(t), N_i(t) \ge n_0\}\right] \end{aligned} \tag{121}$$

The first term can be bounded by $(p_0/N)^{-1} n_0$ from Lemma 24. The rigorous proof can be obtained by the almost same argument as the following analysis of the second term using Theorem 16.

Then, we will bound the second term. Speci cally, we will prove that for $n_0 = \frac{M \log T}{(\epsilon/16)^2}$,

$$E\left[\sum_{t=1}^T 1\{\hat{p}_t \in C_i, A_i^c(t), N_i(t) \ge n_0\}\right] = O(1): \tag{122}$$

First we have

$$\begin{aligned} &E\left[\sum_{t=1}^T 1\{\hat{p}_t \in C_i, A_i^c(t), N_i(t) \ge n_0\}\right] \\ &\le \sum_{m=n_0}^{T} E\left[\sum_{t=1}^T 1\{\hat{p}_t \in C_i, A_i^c(t), N_i(t) = m\}\right]: \end{aligned} \tag{123}$$

Let

$$= \min \quad t : p_t \ 2 \ C_i ; A_i^c(t); N_i(t) = m \quad \wedge (T+1) \tag{124}$$

be the first time such that $p_t \ 2 \ C_i ; A_i^c(t)$ and $N_i(t) = m$ occur. Letting $A_t := p_t \ 2 \ C_i ; A_i^c(t); N_i(t) = m$, $B_t := f i(t) = i g$ and $P_t := p_0 = N$ in Theorem 16, we have

$$E \left[\sum_{t=1}^{T} 1 \ p_t \ 2 \ C_i ; A_i^c(t); N_i(t) = m \right] \quad \frac{N}{p_0} P f \quad T g : \tag{125}$$

Here T implies that

$$\begin{aligned} k\hat{p} \quad p \ k_B &= (\hat{p} \quad p)^> \Big(I + \sum_{j \ 2 [N]} N_j(\ ) S_j^> S_j \Big) A \ (\hat{p} \quad p) \\ & \quad m(\hat{p} \quad p)^> \ S_i^> S_i \ (\hat{p} \quad p) \\ &= m \, k S_i (\hat{p} \quad p) k^2 \quad m(\ =16)^2 ; \end{aligned} \tag{126}$$

where the last inequality follows from Lemma 25. Therefore we have

$$\begin{aligned} E \ \exp(k\hat{p} \quad p \ k_B^2 \ =2) \quad & E \ 1[\quad T] \exp(k\hat{p} \quad p \ k_B^2 \ =2) \\ & \exp(m(\ =16)^2 =2) P f \quad T g : \end{aligned} \tag{127}$$

Note that $j B \ j \ j \ B_T j \quad (1 + T L=M)^M$ for $L = \max_i \sqrt{\operatorname{trace}(S_i^> S_i)} = \max_i k S_i k_F$ by Lemma 10 of Abbasi-Yadkori et al. (2011), where $k \cdot k_F$ is the Frobenius norm. Therefore we have

$$\begin{aligned} E[\exp(k\hat{p} \quad p \ k_B \ =2)] \quad & E \left[\sqrt{\frac{j B \ j}{j B \ j}} \ \exp(k\hat{p} \quad p \ k_B^2 \ =2) \right] \\ & (1 + T L=M)^{M=2} E \left[\frac{\exp(k\hat{p} \quad p \ k_B^2 \ =2)}{\sqrt{j B \ j}} \right] \\ & (1 + T L=M)^{M=2} E \left[\frac{\exp(k\hat{p}_0 \quad p \ k_{B_0}^2 =2)}{\sqrt{j B_0 j}} \right] \qquad (128) \\ = \quad & 1 + \frac{T L}{M} \quad^{M=2} e^{kp \ k^2=2} ; \qquad (129) \end{aligned}$$

where (128) holds since $\frac{\exp(k\hat{p}_t \quad p \ k_{B_t}^2 =2)}{\sqrt{j B_t j}}$ is a supermartingale from Lemma 12. Combining (125), (127), and (129), we obtain

$$\begin{aligned} \sum_{m=n_0}^{1} E \left[\sum_{t=1}^{T} 1 \ p_t \ 2 \ C_i ; A_i^c(t); N_i(t) = m \right] \quad & \frac{N}{p_0} \quad \frac{1}{-} + \frac{T L}{M} \quad^{M=2} e^{kp \ k^2=2} \sum_{m=n_0}^{1} e^{m(\ =16)^2=2} \\ & \frac{N}{p_0} \quad \frac{1}{-} + \frac{T L}{M} \quad^{M=2} e^{kp \ k^2=2} \frac{e^{n_0 \ ^2=2}}{1 \quad e^{(\ =16)^2=2}} : \end{aligned} \tag{130}$$

By choosing $n_0 = \frac{M \log T}{(\ =16)^2}$ we obtain the lemma. □

F.5 Analysis for Case (D)

Lemma 26. For any $i \ 2 \ [N]$,

$$E \left[\sum_{t=1}^{T} h \ 1 \ p_t \ 2 \ C_i ; A_i(t); A_i^c(t) \right]^{i} \quad \frac{48}{9} \frac{M \ + 2}{^2} \frac{N}{p_0} : \tag{131}$$

Remark. To prove the regret upper bound, it is enough to prove Lemma 26 only for $i = 1$. However, for the sake of generality, we prove the lemma for any $i \ 2 \ [N]$.

Before proving Lemma 26, we give two following lemmas.

Lemma 27. For any $i \in [N]$, the event $\tilde{A}_i(t)$ implies $A_i(t)$.

Proof. Using the triangle inequality, we have

$$\begin{aligned} \|S_i p - S_i \hat{p}_t\| - \|S_i p - S_i p_t\| &\leq \|S_i p_t - S_i \hat{p}_t\| \\ &= 8 + \|S_i\| \frac{}{16 \max_i \|S_i\|} \leq 4; \end{aligned} \tag{132}$$

which completes the proof. □

Now, Lemma 26 can be intuitively proven because from Lemma 27, $\tilde{A}_i(t)$ implies $A_i(t)$, and the events $A_i(t)$ and $A_i^c(t)$ does not simultaneously occur many times.

Let $t = \{\tau_1, \ldots, \tau_m\}$ be the time of the first m times that the event $\{p_t \in C_i, A_i(t), N_i(t) = n_i\}$ occurred (not $\{p_t \in C_i, A_i(t), N_i(t) = n_i\}$). In other words, we define

τ_1 : the first time that $p_t \in C_i, A_i(t)$ and $N_i(t) = n_i$ occurred

τ_2 : the second time that $p_t \in C_i, A_i(t)$ and $N_i(t) = n_i$ occurred

. . . .

Now we prove the following lemma using Lemma 9.

Lemma 28. For any $0 < 1=2$,

$$P\left\{A_i^c(\tau_k) \mid A_i(\tau_k), \tau_k = t\right\} \leq \exp\left(\frac{9}{16}\right)^2 n_i (1 - 2)^{M=2}. \tag{133}$$

Proof. Recall that $T_i = \{p \in R^M : \|S_i p - S_i p\| > \}$. We follow a similar argument as the analysis for Lemma 19. Since $p_t \sim N(B_t^{-1} b_t, B_t^{-1})$, the squared Mahalanobis distance $\|B_t^{1=2}(p - \hat{p}_t)\|^2$ follows the chi-squared distribution with M degree of freedom. Hence, for $f_M(a) = P\{X \leq a\}$ $\{X \sim \chi^2_M\}$, we have

$$P\left\{A_i^c(\tau_k) \mid A_{i;n_i}, \tau_k = t\right\} \leq h\left(\inf_{p \in T_i} \|B_t^{1=2}(p - \hat{p}_t)\|^2\right). \tag{134}$$

Then, Eq. (133) directly follows from Lemma 9. □

Proof of Lemma 26. From Lemma 27, the event $\tilde{A}_i(t)$ implies $A_i(t)$. Hence, it is enough to derive the upper bound for

$$E\left[\sum_{t=1}^{T} h\left(1 - p_t \in C_i, A_i(t), A_i^c(t)\right)\right] \tag{135}$$

instead of the bound for

$$E\left[\sum_{t=1}^{T} h\left(1 - p_t \in C_i, A_i(t), A_i^c(t)\right)\right]. \tag{136}$$

Using Lemma 28, we can bound the term for case (D) from above as

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\hat{p}_t \in C_i, A_i(t), A_i^c(t)\right]\right] \\
&= \mathbb{E}\left[\sum_{n_i=1}^{T}\sum_{t=1}^{T} \mathbb{1}\left[A_i(t), A_i^c(t), N_i(t) = n_i\right]\right] \\
&= \sum_{n_i=1}^{T}\sum_{t=1}^{T} \mathbb{P}\left\{A_i(t), A_i^c(t), N_i(t) = n_i\right\} \\
&= \sum_{n_i=1}^{T}\sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\left\{A_i(t), A_i^c(t), \tau_k = t\right\} \quad (\text{the event } \{\tau_k = t\} \text{ is exclusive for fixed } n_i) \\
&= \sum_{n_i=1}^{T}\sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{A_i(t), \tau_k = t\}\,\mathbb{P}\left\{A_i^c(t) \mid A_i(t), \tau_k = t\right\} \\
&\le \sum_{n_i=1}^{T}\sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{A_i(t), \tau_k = t\}\, C e^{-\epsilon n_i} \quad (\text{by Lemma 28}) \\
&= \sum_{n_i=1}^{T} C e^{-\epsilon n_i} \sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{A_i(t), \tau_k = t\} \\
&\le \sum_{n_i=1}^{T} C e^{-\epsilon n_i} \sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{\tau_k = t\} \\
&\le \sum_{n_i=1}^{T} C e^{-\epsilon n_i} \sum_{k=1}^{T} \mathbb{P}\{\tau_k \text{ exists}\} \\
&\le \sum_{n_i=1}^{T} C e^{-\epsilon n_i} \sum_{k=1}^{T} \left(1 - \frac{p_0}{N}\right)^{k-1} \quad (\text{by } \hat{p}_{\tau_s} \notin C_i \text{ for } s = 1, \ldots, k-1) \\
&\le 3C \frac{1}{e^{\epsilon} - 1} \frac{N}{p_0} \\
&\le \frac{48}{9} \frac{M+2}{\delta^2} \frac{N}{p_0},
\end{aligned} \tag{137}
$$

where $\epsilon = \frac{9\delta^2}{16}$, $C = (1 - 2\delta)^{-\frac{M}{2}}$, and in the last inequality we select the optimal δ and use $1 + x \le e^x$. □

F.6 Analysis for Case (E)

Lemma 29. For any $i \neq 1$,

$$
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\hat{p}_t \in C_i, A_i(t)\right]\right] \le \frac{2^{5M/2+7}(\lambda M/2 + 1)e^{\lambda^2\|p^*\|^2/2}}{\Delta^{2\lambda M+2}\lambda^{M/2+1}}, \tag{138}
$$

where Δ is defined in (26) and satisfies $\Delta \le \min_{p \in C_1^c} \|p - p^*\|/3$, and

$$
\Delta := \min_{\hat{p}:(L_1 - L_i)^{\top}\hat{p} \ge 0;\ \|S_i(\hat{p} - p^*)\| \le \Delta/8} \|S_1(\hat{p} - p^*)\|. \tag{139}
$$

We prove Lemma 29 using Lemma 12 and Theorem 16.

Remark. The upper bound in (138) goes to infinite when we set $\lambda = 0$, that is, a flat prior is used. However, this is not the essential effect of the prior but just comes from the minimum eigenvalue of B_1. In fact, we can see from the proof that a similar bound can be obtained for $\lambda = 0$ if we run some deterministic initialization until B_t becomes positive definite.

Proof. We evaluate each term in the summation using Theorem 16 with

$$A_t = \{p_t \in C_i, \|S_i(p_t - p)\| = 8, N_1(t) = n\},$$
$$B_t = \{p_t \in C_1\}. \tag{140}$$

for $n \in [T]$. Recall that

$$g_t(p) = \frac{1}{\sqrt{(2\pi)^M |B_t^{-1}|}} \exp\left(-\frac{1}{2}\|p - \hat{p}_t\|^2_{B_t}\right) \tag{141}$$

is the probability density function of p_t given $F_t = \{B_t, b_t\}$. Using δ defined in (80), it holds for any $\epsilon \in [T]$ that

$$\begin{aligned}
P\{B_\epsilon \mid F_\epsilon\} &= P\{p_\epsilon \in C_1 \mid F_\epsilon\} \\
&= \int_{p \in C_1} g_\epsilon(p)dp \\
&\le \int_{p:\|p-p_k\| \le 3\delta} g_\epsilon(p)dp \\
&\le \sup_{p:\|p-p_k\| \le 2\delta} \int_{p':\|p'-p\| \le \delta} g_\epsilon(p')dp' \\
&\le \sup_{p:\|p-p_k\| \le 2\delta} \inf_{p':\|p'-p\| \le \delta} g_\epsilon(p')\mathrm{Vol}(\{p'': \|p''-p\| \le \delta\}) \\
&= \frac{(\sqrt{\pi}\delta)^M}{\Gamma(M/2+1)} \sup_{p:\|p-p_k\| \le 2\delta} \inf_{p':\|p'-p\| \le \delta} g_\epsilon(p') \\
&= \frac{(\sqrt{\pi}\delta)^M \sqrt{|B_\epsilon|}}{\Gamma(M/2+1)} \exp\left(\frac{1}{2} \inf_{p:\|p-p_k\| \le 2\delta} \sup_{p':\|p'-p\| \le \delta} \|p'-\hat{p}_\epsilon\|^2_{B_\epsilon}\right) \\
&\le \frac{(\sqrt{\pi}\delta)^M \sqrt{|B_\epsilon|}}{\Gamma(M/2+1)} \exp\left(\frac{\|p_k-\hat{p}_\epsilon\|^2_{B_\epsilon} + \delta\sqrt{n}}{2}\right);
\end{aligned} \tag{142}$$

$$= \frac{(\sqrt{\pi}\delta)^M \sqrt{|B_\epsilon|}}{\Gamma(M/2+1)} \exp\left(\frac{\|p_k-\hat{p}_\epsilon\|^2_{B_\epsilon} + \delta\sqrt{n}}{2}\right); \tag{143}$$

where (142) follows since $\{p : \|p - p_k\| \le 3\delta\} \subset \{p' : \|p' - p_0\| \le \delta\}$ for any p_0 such that $\|p_0 - p_k\| \le 2\delta$, and the last inequality follows from Theorem 15. To apply Theorem 15, we used Lemma 27.

Now we define a stochastic process corresponds to (143) as

$$P_t = \frac{(\sqrt{\pi}\delta)^M \sqrt{|B_t|}}{\Gamma(M/2+1)} \exp\left(\frac{\|p_k-\hat{p}_t\|^2_{B_t} + \delta\sqrt{n}}{2}\right). \tag{144}$$

Then, by Lemma 12,

$$\begin{aligned}
E[P_{t+1}^{-1} \mid F_t] &\le \frac{\Gamma(M/2+1)}{(\sqrt{\pi}\delta)^M} e^{\delta\sqrt{n}/2} E\left[\sqrt{|B_{t+1}|}\, E\left[\exp\left(\frac{\|p_k-\hat{p}_t\|^2_{B_{t+1}}}{2}\right) \mid F_t, S_{i(t)}\right] \mid F_t\right] \\
&\le \frac{\Gamma(M/2+1)}{(\sqrt{\pi}\delta)^M} e^{\delta\sqrt{n}/2} E\left[\sqrt{|B_t|}\exp\left(\frac{\|p_k-\hat{p}_t\|^2_{B_t}}{2}\right) \mid F_t\right] \\
&= P_t^{-1};
\end{aligned} \tag{145}$$

which means that P_t^{-1} is a supermartingale. Therefore we can apply Theorem 16 and obtain

$$\begin{aligned}
E\left[\sum_{t=1}^{T} 1[\hat{p}_t \in C_i, \|S_i(p_t - p)\| = 8, N_1(t) = n]\right] &\le E[1[\cdot \le T]P^{-1}] \\
&\le E[P^{-1}] \\
&\le E[P_1^{-1}] \\
&= \frac{\Gamma(M/2+1)e^{\delta\|p_k-\hat{p}\|^2/2}}{(\sqrt{\pi}\delta)^M} e^{\delta\sqrt{n}/2}.
\end{aligned} \tag{146}$$

Finally we have

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T} 1[p_t \in C_i, \|S_i(p_t - \bar{p})\|^2 = 8]\right] \\
&= \sum_{n=1}^{\infty} E\left[\sum_{t=1}^{T} 1[p_t \in C_i, \|S_i(p_t - \bar{p})\|^2 = 8, N_1(t) = n]\right] \\
&\quad \frac{(M/2+1)e^{-2\pi\|k\|^2/2}}{(\sqrt{2\pi}/2)^M} \sum_{n=1}^{\infty} e^{-\sqrt{n}/2} \\
&\quad \frac{(M/2+1)e^{-2\pi\|k\|^2/2}}{(\sqrt{2\pi}/2)^M} \int_0^{\infty} e^{-\sqrt{x}/2} dx \\
&= \frac{(M/2+1)e^{-2\pi\|k\|^2/2}}{(\sqrt{2\pi}/2)^M} \frac{2}{(\sqrt{2\pi}/2)^2} \quad (2) \\
&= \frac{2^{M/2+3}(M/2+1)e^{-2\pi\|k\|^2/2}}{2^{M+2}\pi^{M/2+1}}; \qquad (147)
\end{aligned}
$$

which completes the proof. □

G Property of Dynamic Pricing Games

In this appendix, we will see a property of dp-easy games.

Proposition 30. Consider any dp-easy games with $c > 1$. Then, any two actions in the game are neighbors.

Remark. In section 5, we considered dp-easy games with $c > 0$, but this can be relaxed to $c > 1$ to prove Proposition 30.

Proof. Take any two different actions $j, k \in [N]$ such that $j < k$. From the definition of the loss matrix in dp-easy games, we have $e_j \in C_j$ and $e_k \in C_k$.

First, we will find $\alpha \in [0, 1]$ such that

$$\alpha e_j + (1-\alpha)e_k \in C_j \cap C_k. \qquad (148)$$

From the definition of the loss matrix, the ith element of $L(\alpha e_j + (1-\alpha)e_k) \in P_M$ is

$$\begin{cases} \alpha c + (1-\alpha)(i) & (1 \le i \le j) \\ (j+1 \le i \le k) \\ c & (k < i \le N) \end{cases}. \qquad (149)$$

It is easy to see that the indices which give the minimum value of (149) is j or k. Thus, to achieve the condition (148), the following should be satisfied,

$$\alpha c + (1-\alpha)(k) = j, \qquad (150)$$

which is equivalent to

$$\alpha = \frac{k-j}{c+k} (=: \theta). \qquad (151)$$

Note that we have $0 \le \theta \le 1$ for any $c > 1$.

Next, we introduce the following definitions.

$$p^{(j,k)} := \alpha e_j + (1-\alpha)e_k \in C_j \cap C_k, \qquad (152)$$

$$\mathrm{Ball}^{(j,k)} := \{p \in P_M : \|p - p^{(j,k)}\| \le \}, \qquad (153)$$

$$L^{(x)} := L(p^{(j,k)} + x) \in \mathbb{R}^N. \qquad (154)$$

To prove the proposition, it is enough to prove the following: there exists $\epsilon > 0$, $\mathrm{Ball}^{(j,k)}(\epsilon) \subset C_j \cup C_k$.

To prove this, it is enough to prove that, there exists $\epsilon > 0$,

$$\min_{x \in \mathbb{R}^M : \|x\| = 1} \max_{i \in [N] \setminus \{j,k\}} (L^{(x)})_i - (L^{(x)})_j - (L^{(x)})_i + (L^{(x)})_k > 0. \tag{155}$$

We will prove (155) in the following. Take any $i \in [N] \setminus \{j,k\}$ and

$$\epsilon := \min_{i:1 \le i < j} \frac{1}{2} \frac{\delta_j - \delta_i}{\|L_j - L_i\|} \wedge \min_{i:j < i < k} \frac{1}{2} \frac{(1-\delta)(\delta_k - \delta_i)}{\|L_k - L_i\|} \wedge \min_{i:k < i \le N} \frac{1}{2} \frac{c + \delta_j - \delta_i}{\|L_j - L_i\|}. \tag{156}$$

Note that the δ used here is different from the one used in the proof of the regret upper bounds.

Case (A): When $1 \le i < j$, using Cauchy Schwarz inequality, we have

$$\begin{aligned}
&(L^{(x)})_i - (L^{(x)})_j - (L^{(x)})_i + (L^{(x)})_k - (L^{(x)})_i + (L^{(x)})_j \\
&= (\delta_i + L_i^\top x) - (\delta_j + L_j^\top x) \\
&= (\delta_j - \delta_i) - (L_j - L_i)^\top x \\
&\ge (\delta_j - \delta_i) - \|L_j - L_k\| \|x\| \\
&\ge (\delta_j - \delta_i) - \|L_j - L_i\| \\
&\ge \frac{1}{2}(\delta_j - \delta_i) \\
&> 0.
\end{aligned} \tag{157}$$

The arguments for cases (B) and (C) follow in the similar manner as case (A).

Case (B): When $j < i < k$, we have

$$\begin{aligned}
&(L^{(x)})_i - (L^{(x)})_j - (L^{(x)})_i + (L^{(x)})_k - (L^{(x)})_i + (L^{(x)})_k \\
&= -c + (1-\delta)(\delta_i) + L_i^\top x - c + (1-\delta)(\delta_k) + L_k^\top x \\
&= (1-\delta)(\delta_k - \delta_i) - (L_i - L_k)^\top x \\
&\ge (1-\delta)(\delta_k - \delta_i) - \|L_i - L_k\| \\
&\ge \frac{1}{2}(1-\delta)(\delta_k - \delta_i) \\
&> 0.
\end{aligned} \tag{158}$$

Case (C) When $k < i \le N$, we have

$$\begin{aligned}
&(L^{(x)})_i - (L^{(x)})_j - (L^{(x)})_i + (L^{(x)})_k - (L^{(x)})_i + (L^{(x)})_j \\
&= (-c + L_i^\top x) - (\delta_j + L_j^\top x) \\
&\ge -c + \delta_j - \|L_j - L_i\| \|x\| \\
&\ge -c + \delta_j - \|L_j - L_i\| \\
&\ge \frac{1}{2}(-c + \delta_j) \\
&> 0.
\end{aligned} \tag{159}$$

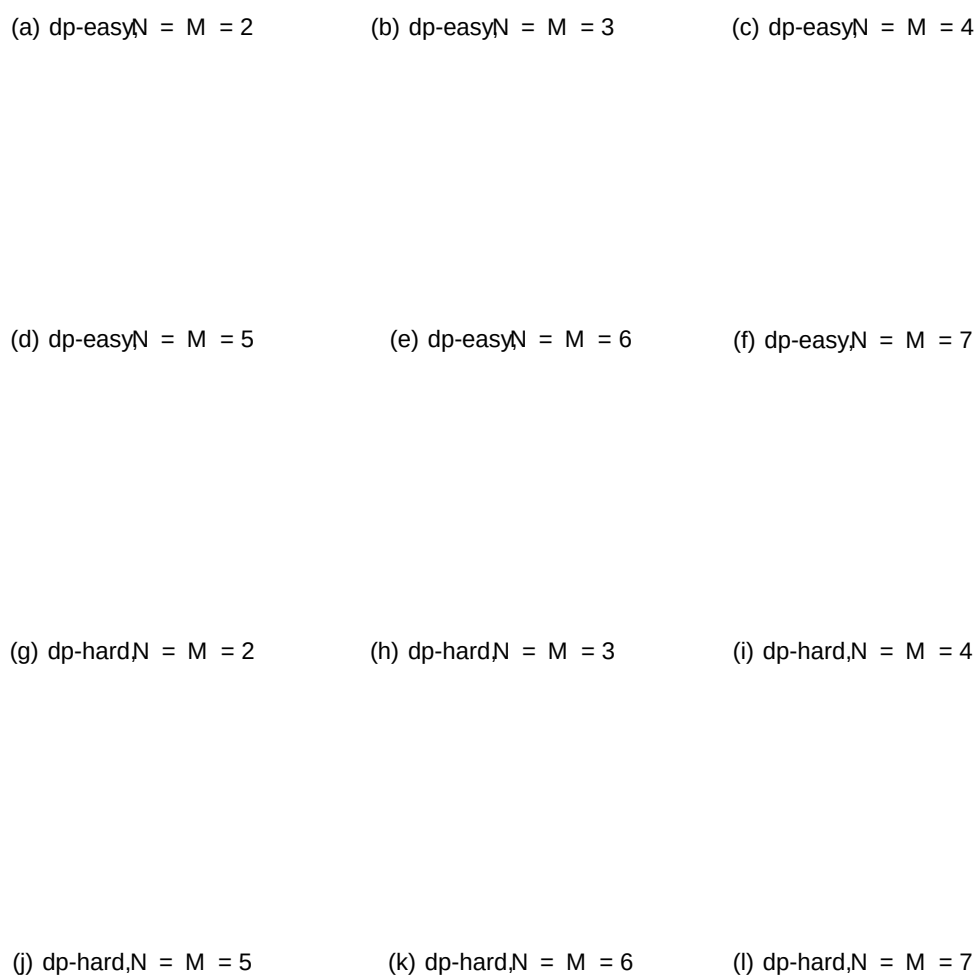
Summing up the argument for cases (A) to (C), the proof is completed. □

H Details and Additional Results of Experiments

Here we give the specific values of the opponent's strategy used in Section 5 and show the extended experimental results for performance comparison. Table 2 summarizes the values of opponent's strategy used in this appendix and Section 5. Figure 3 shows the empirical comparison of the proposed algorithms against the benchmark methods, and Figure 4 shows the number of the rejected times. We can see the same tendency as Section 5, that is, TSPM performs the best and the number of rejections does not increase with the time step t.

Table 2: The values of the opponent's strategy.

# of outcomes M	opponent's strategy p
2	[0.7, 0.3]
3	[0.5, 0.3, 0.2]
4	[0.3, 0.3, 0.3, 0.1]
5	[0.2, 0.3, 0.3, 0.1, 0.1]
6	[0.2, 0.2, 0.3, 0.1, 0.1, 0.1]
7	[0.2, 0.2, 0.3, 0.1, 0.1, 0.05, 0.05]



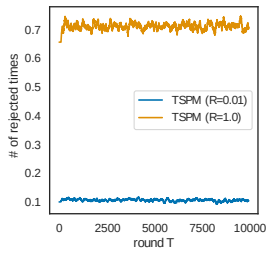
(a) dp-easy, $N = M = 2$ (b) dp-easy, $N = M = 3$ (c) dp-easy, $N = M = 4$

(d) dp-easy, $N = M = 5$ (e) dp-easy, $N = M = 6$ (f) dp-easy, $N = M = 7$

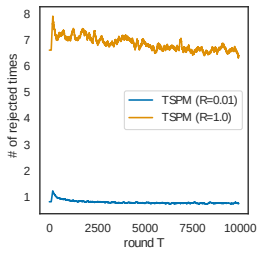
(g) dp-hard, $N = M = 2$ (h) dp-hard, $N = M = 3$ (i) dp-hard, $N = M = 4$

(j) dp-hard, $N = M = 5$ (k) dp-hard, $N = M = 6$ (l) dp-hard, $N = M = 7$

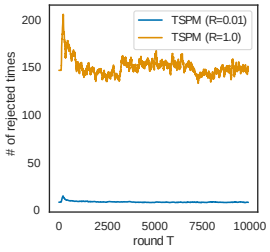
Figure 3: Regret-round plots of the algorithms. The solid lines indicate the average over 100 independent trials. The thin fillings are the standard error.



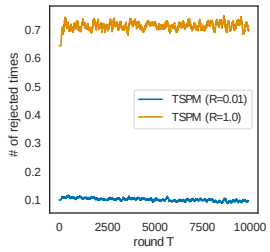
(a) dp-easy, $N = M = 2$



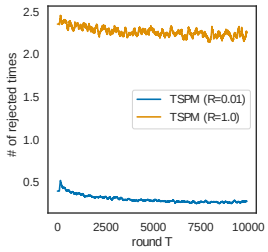
(c) dp-easy, $N = M = 4$



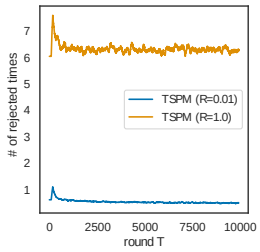
(e) dp-easy, $N = M = 6$



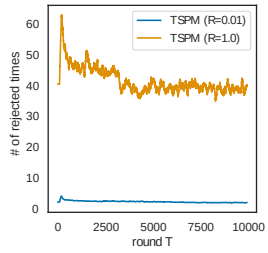
(g) dp-hard, $N = M = 2$



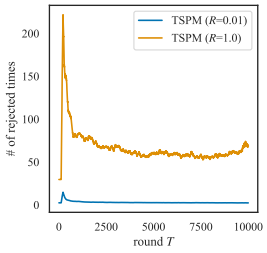
(h) dp-hard, $N = M = 3$



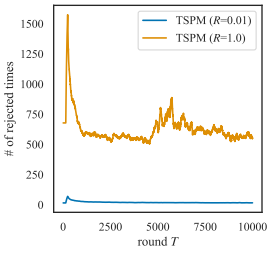
(i) dp-hard, $N = M = 4$



(j) dp-hard, $N = M = 5$



(k) dp-hard, $N = M = 6$



(l) dp-hard, $N = M = 7$

Figure 4: The number of rejected times by the accept-reject sampling. The solid lines indicate the average over 100 independent trials after taking moving average with window size 100.