Proceedings of Machine Learning Research vol 117:1–39, 2020 31st International Conference on Algorithmic Learning Theory

First-Order Bayesian Regret Analysis of Thompson Sampling

Sébastien Bubeck SEBUBECK@MICROSOFT.COM
Microsoft Research AI
Redmond, WA 98052, USA

Mark Sellke MSELLKE@STANFORD.EDU
Department of Mathematics
Stanford University
Stanford, CA 94305, USA

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

We address online combinatorial optimization when the player has a prior over the adversary's sequence of losses. In this framework, Russo and Van Roy proposed an information-theoretic analysis of Thompson Sampling based on the *information ratio*, resulting in optimal worst-case regret bounds. In this paper we introduce three novel ideas to this line of work. First we propose a new quantity, the *scale-sensitive information ratio*, which allows us to obtain more refined *first-order regret bounds* (i.e., bounds of the form $O(\sqrt{L^*})$ where L^* is the loss of the best combinatorial action). Second we replace the entropy over combinatorial actions by a *coordinate entropy*, which allows us to obtain the first optimal worst-case bound for Thompson Sampling in the combinatorial setting. We additionally introduce a novel link between Bayesian agents and frequentist confidence intervals. Combining these ideas we show that the classical multi-armed bandit first-order regret bound $\widetilde{O}(\sqrt{dL^*})$ still holds true in the more challenging and more general semi-bandit scenario. This latter result improves the previous state of the art bound $\widetilde{O}(\sqrt{(d+m^3)L^*})$ by Lykouris, Sridharan and Tardos. We tighten these results by leveraging a recent insight of Zimmert and Lattimore connecting Thompson Sampling and online stochastic mirror descent, which allows us to replace the Shannon entropy with more general mirror maps.

Keywords: Multi-armed bandit, Thompson Sampling.

1. Introduction

We first recall the general setting of online combinatorial optimization with both full feedback (full information game) and limited feedback (semi-bandit game). Let $\mathcal{A} \subset \{0,1\}^d$ be a fixed set of *combinatorial actions*, and assume that $m = \|a\|_1$ for all $a \in \mathcal{A}$. An (oblivious) adversary selects a sequence $\ell_1, \dots, \ell_T \in [0,1]^d$ of linear functions, without revealing it to the player. At each time step $t = 1, \dots, T$, the player selects an action $a_t \in \mathcal{A}$, and suffers the instantaneous loss $\langle \ell_t, a_t \rangle$. The following feedback on the loss function ℓ_t is then obtained: in the full information game the entire loss vector ℓ_t is observed, and in the semi-bandit game only the loss on active coordinates is observed (i.e., one observes $\ell_t \odot a_t$ where $\odot$ denotes the entrywise product). Importantly the player has access to external randomness, and can select their action a_t based on the observed feedback so far. The player's objective is to minimize its total expected loss $L_T = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, a_t \rangle\right]$. The player's perfomance at the end of the game is measured through the *regret* R_T, which is the difference between the achieved cumulative loss L_T and the best one could have done with a fixed action. That is, with $L^* = \min_{a \in \mathcal{A}} \sum_{t=1}^T \langle \ell_t, a \rangle$, one has $R_T = L_T - L^*$. The optimal worst-case regret ($\sup_{\ell_1,\dots,\ell_T \in [0,1]^d} R_T$) is known for both the full information and semi-bandit game. It is respectively of order $m\sqrt{T}$ ((KWK10)) and $\sqrt{mdT}$ ((ABL14)).

1.1. First-order regret bounds

It is natural to hope for strategies with regret $R_T = o(L^*)$, for one can then claim that $L_T = (1 + o(1))L^*$ (in other words the player's performance is close to the optimal in-hindsight performance up to a smaller order term). However, worst-case bounds fail to capture this behavior when $L^* \ll T$. The concept of *first-order regret bound* tries to remedy this issue, by asking for regret bounds scaling with L^* instead of T. In (KWK10) an optimal version of such a bound is obtained for the full information game:

Theorem 1 ((KWK10)) *In the full information game, there exists an algorithm such that for any loss sequence one has $R_T = \widetilde{O}(\sqrt{m\mathbb{E}[L^*]})$.*

The state of the art for first-order regret bounds in the semi-bandit game is more complicated. It is known since (AAGO06) that for $m = 1$ (i.e., the famous multi-armed bandit game) one can have an algorithm with regret $R_T = \widetilde{O}(\sqrt{dL^*})$. On the other hand for $m > 1$ the best bound due to (LST18) is $\widetilde{O}(\sqrt{(d + m^3)L^*})$. A byproduct of our main result (Theorem 4 below) is to give the first optimal first-order regret bound for the semi-bandit game[1]:

Theorem 2 *In the semi-bandit game, there exists an algorithm such that for any loss sequence one has $R_T = \widetilde{O}(\sqrt{d\mathbb{E}[L^*]})$.*

The above bound is optimal because the minimax regret in this setting is $\widetilde{\Theta}(\sqrt{mdT})$ ((ABL14)) and L^* can be as large as mT. We derive this result[2] using the recipe first proposed (in the context of partial feedback) in (BDKP15). Namely, to show the existence of a randomized strategy with regret bounded by B_T for any loss sequence, it is sufficient to show that for any *distribution* over loss sequences there exists a strategy with regret bounded by B_T in expectation. Indeed, this equivalence is a simple consequence of the Sion minimax theorem (BDKP15). In other words to prove Theorem 2 it is sufficient to restrict our attention to the *Bayesian scenario*, where one is given a prior distribution ν over the loss sequence $(\ell_1, \dots, \ell_T) \in [0,1]^{[d]\times[T]}$ and aims for small expected regret with respect to that prior. Importantly note that there is no independence whatsoever in such a random loss sequence, either across times or across coordinates for a fixed time. Rather, the prior is completely arbitrary over the Td different values $\ell_t(i)$.

The rest of the paper is dedicated to the (first-order) regret analysis of a particular Bayesian strategy, the famous Thompson Sampling ((Tho33)). In particular we will show that Thompson Sampling achieves both the bounds of Theorem 1 and Theorem 2.

1.2. Thompson Sampling

In the Bayesian setting one has access to a *prior distribution* on the optimal action

$$a^* = \operatorname*{argmax}_{a \in \mathcal{A}} \sum_{t=1}^{T} \langle \ell_t, a \rangle.$$

1. By $\widetilde{O}(\cdot)$ we suppress logarithmic terms, even $\log(T)$. However all our bounds stated in the main body state explicitly the logarithmic dependency.
2. In fact this bound can also be obtained more directly with mirror descent and an entropic regularizer as in (ABL14).

In particular, one can update this distribution as more observations on the loss sequence are collected. More precisely, denote p_t for the posterior distribution of a^* given all the information at the beginning of round t (i.e., in the full information this is $\ell_1, \dots, \ell_{t-1}$ while in semi-bandit it is $\ell_1 \odot a_1, \dots, \ell_{t-1} \odot a_{t-1}$). Then Thompson Sampling simply plays an action a_t at random from p_t.

This strategy has recently regained interest, as it is both efficient and successful in practice for simple priors ((CL11)) and particularly elegant in theory. A breakthrough in the understanding of Thompson Sampling's regret was made in (RVR16) where an information theoretic analysis was proposed. They consider in particular the combinatorial setting for which they prove the following result:

Theorem 3 ((RVR16)) *Assume that the prior ν is such that the sequence $(\ell_1, \dots, \ell_T)$ is i.i.d. Then in the full information game Thompson Sampling satisfies $\mathbb{E}[R_T] = \widetilde{O}(m^{3/2}\sqrt{T})$, and in the semi-bandit game it satisfies $\mathbb{E}[R_T] = \widetilde{O}(m\sqrt{dT})$.*

Assume furthermore that the prior ν is such that, for any t, conditionally on $\ell_1, \dots, \ell_{t-1}$ one has that $\ell_t(1), \dots, \ell_t(d)$ are independent. Then Thompson Sampling satisfies respectively $\mathbb{E}^\nu[R_T] = \widetilde{O}(m\sqrt{T})$ and $\mathbb{E}^\nu[R_T] = \widetilde{O}(\sqrt{mdT})$ in the full information and semi-bandit game.

It was observed in (BDKP15) that the assumption of independence across times is immaterial in the information theoretic analysis of Russo and Van Roy. However it turns out that the independence across coordinates (conditionally on the history) in Theorem 3 is key to obtain the worst-case optimal bounds $m\sqrt{T}$ and $\sqrt{mdT}$. One of the contributions of our work is to show how to appropriately modify the notion of entropy to remove this assumption.

Most importantly, we propose a new analysis of Thompson Sampling that allows us to prove *first-order regret bounds*. In various forms we show the following result:

Theorem 4 *For any prior ν, Thompson Sampling satisfies in the full information game $\mathbb{E}^\nu[R_T] = \widetilde{O}(\sqrt{m\mathbb{E}[L^*]})$. Furthermore in the semi-bandit game, $\mathbb{E}^\nu[R_T] = \widetilde{O}(\sqrt{d\mathbb{E}[L^*]})$.*

To the best of our knowledge such guarantees were not known for Thompson Sampling even in the full-information case with $m = 1$ (the so-called expert setting of (CBFH$^+$97)). Our analysis can be combined with recent work in (ZL19) which allows for improved estimates based on using mirror maps besides the Shannon entropy.

Finally, we note that Thompson sampling against certain artificial prior distributions is also known to obey frequentist regret bounds in the stochastic case ((AG12; LTW20)). However we emphasize that in this paper, Thompson Sampling assumes access to the true prior distribution for the loss sequence and the guarantees are for expected Bayesian regret with respect to that prior.

2. Information ratio and scale-sensitive information ratio

As a warm-up, and to showcase one of our key contributions, we focus here on the full information case with $m = 1$ (i.e., the expert setting). We start by recalling the general setting of Russo and Van Roy's analysis (Subsection 2.1), and how it applies in this expert setting (Subsection 2.2). We then introduce a new quantity, the scale-sensitive information ratio, and show that it naturally implies a first-order regret bound (Subsection 2.3). We conclude this section by showing a new bound between two classical distances on distributions (essentially the chi-squared and the relative entropy), and we explain how to apply it to control the scale-sensitive information ratio (Subsection 2.4).

2.1. Russo and Van Roy's analysis

Let us denote $X_t \in \mathbb{R}^d$ for the feedback received at the end of round t. That is in full information one has $X_t = \ell_t$, while in semi-bandit one has $X_t = \ell_t \odot a_t$. Let us denote $\mathbb{P}_t$ for the posterior distribution of $\ell_1, \ldots, \ell_T$ conditionally on $a_1, X_1, \ldots, a_{t-1}, X_{t-1}$. We write $\mathbb{E}_t$ for the integration with respect to $\mathbb{P}_t$ and $a_t \sim p_t$ (recall that p_t is the distribution of a^* under $\mathbb{P}_t$). Let I_t be the mutual information, *under the posterior distribution* $\mathbb{P}_t$, between a^* and X_t, that is $I_t = H(p_t) - \mathbb{E}_t[H(p_{t+1})]$. Let $r_t = \mathbb{E}_t[\langle \ell_t, a_t - a^* \rangle]$ be the instantaneous regret at time t. The information ratio introduced by Russo and Van Roy is defined as:

$$\Gamma_t := \frac{r_t^2}{I_t} \,. \tag{1}$$

The point of the information ratio is the following result:

Proposition 5 (Proposition 1, (RVR16)) *Consider a strategy such that $\Gamma_t \leq \Gamma$ for all t. Then one has*

$$\mathbb{E}[R_T] \leq \sqrt{T \cdot \Gamma \cdot H(p_1)} \,,$$

where $H(p_1)$ denotes the Shannon entropy of the prior distribution p_1 (in particular $H(p_1) \leq \log(d)$).

Proof The main calculation is as follows:

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^T r_t\right] \leq \sqrt{T \cdot \mathbb{E}\left[\sum_{t=1}^T r_t^2\right]} \leq \sqrt{T \cdot \Gamma \cdot \mathbb{E}\left[\sum_{t=1}^T I_t\right]} \,. \tag{2}$$

Moreover it turns out that the total information accumulation $\mathbb{E}\left[\sum_{t=1}^T I_t\right]$ can be easily bounded, by simply observing that the mutual information can be written as a drop in entropy, yielding the bound:

$$\mathbb{E}\left[\sum_{t=1}^T I_t\right] \leq H(p_1) \,.$$

■

2.2. Pinsker's inequality and Thompson Sampling's information ratio

We now describe how to control the information ratio (1) of Thompson Sampling in the expert setting. First note that the posterior distribution p_t of $a^* \in \{e_1, \ldots, e_d\}$ satisfies (with a slight abuse of notation by viewing p_t as a vector in $\mathbb{R}^d$): $p_t = \mathbb{E}_t[a^*]$. In particular this means that:

$$\begin{aligned} r_t = \mathbb{E}_t[\langle \ell_t, a_t - a^* \rangle] = \mathbb{E}_t\left[\mathbb{E}_{t+1}[\langle \ell_t, a_t - a^* \rangle]\right] = \\ \mathbb{E}_t\left[\langle \ell_t, \mathbb{E}_{t+1}[(a_t - a^*)] \rangle\right] = \mathbb{E}_t[\langle \ell_t, p_t - p_{t+1} \rangle] \leq \frac{1}{2}\mathbb{E}_t[\|p_t - p_{t+1}\|_1] \end{aligned} \tag{3}$$

where the inequality uses that $\|\ell_t - (\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})\|_\infty \leq \frac{1}{2}$. Now combining (3) with Jensen followed by Pinsker's inequality yields:

$$r_t^2 \leq \frac{1}{2} \cdot \mathbb{E}_t[\mathrm{Ent}(p_{t+1}, p_t)] \,,$$

where we denote $\mathrm{Ent}(p,q) = \sum_{i=1}^d p(i)\log(p(i)/q(i))$ (recall that Pinsker's inequality is simply $\|p - q\|_1^2 \leq 2 \cdot \mathrm{Ent}(p,q)$). Furthermore classical rewriting of the mutual information shows that the quantity $\mathbb{E}_t[\mathrm{Ent}(p_{t+1}, p_t)]$ is equal to I_t (see [Proposition 4, (RVR16)] for more details). In other words $r_t^2 \leq \frac{I_t}{2}$ and thus:

Lemma 6 ((RVR16)) *In the expert setting, Thompson Samping's information ratio (1) satisfies $\Gamma_t \leq \frac{1}{2}$ for all t.*

Using Lemma 6 in Proposition 5 one obtains the following worst case optimal regret bound for Thompson Sampling in the expert setting:

$$\mathbb{E}[R_T] \leq \sqrt{\frac{T\log(d)}{2}}\,.$$

2.3. Scale-sensitive information ratio

The information ratio (1) was designed to derive $\sqrt{T}$-type bounds (see Proposition 5). To obtain $\sqrt{L^*}$-type regret we propose the following quantity which we coin the *scale-sensitive information ratio*:

$$\Lambda_t := \frac{(r_t^+)^2}{I_t \cdot \mathbb{E}_t[\langle \ell_t, a_t\rangle]}\,, \tag{4}$$

where $r_t^+ := \mathbb{E}_t[\langle \ell_t, ReLU(p_t - p_{t+1})\rangle]$. With this new quantity we obtain the following refinement of Proposition 5:

Proposition 7 *Consider a strategy such that $\Lambda_t \leq \Lambda$ for all t. Then one has*

$$\mathbb{E}[R_T] \leq \sqrt{\mathbb{E}[L^*]\cdot\Lambda\cdot H(p_1)} + \Lambda\cdot H(p_1)\,.$$

Proof The main calculation is as follows:

$$\begin{aligned}
\mathbb{E}[R_T] \leq \mathbb{E}\left[\sum_{t=1}^T r_t^+\right] \quad &\leq \quad \sqrt{\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t[\langle \ell_t, a_t\rangle]\right]\cdot\mathbb{E}\left[\sum_{t=1}^T \frac{(r_t^+)^2}{\mathbb{E}_t[\langle \ell_t, a_t\rangle]}\right]} \\
&\leq \quad \sqrt{\mathbb{E}[L_T]\cdot\Lambda\cdot\mathbb{E}\left[\sum_{t=1}^T I_t\right]} \\
&\leq \quad \sqrt{\mathbb{E}[L_T]\cdot\Lambda\cdot H(p_1)}\,.
\end{aligned}$$

It only remains to use the fact that $a - b \leq \sqrt{ac}$ implies that $a - b \leq \sqrt{bc} + c$ for non-negative a, b, c. ■

2.4. Reversed chi-squared/relative entropy inequality

We now describe how to control the scale-sensitive information ratio (4) of Thompson Sampling in the expert setting. As we saw in Subsection 2.2, the two key inequalites in the Russo-Van Roy information ratio analysis are a simple Cauchy-Schwarz followed by Pinsker's inequality (recall (3)):

$$r_t = \mathbb{E}_t[\langle \ell_t, p_t - p_{t+1}\rangle] \leq \mathbb{E}_t[\|\ell_t\|_\infty \cdot \|p_t - p_{t+1}\|_1] \leq \sqrt{\mathbb{E}_t[\mathrm{Ent}(p_{t+1}, p_t)]} = \sqrt{I_t}\,.$$

In particular, as far as first-order regret bounds are concerned, the "scale" of the loss ℓ_t is lost in the first Cauchy-Schwarz. To control the scale-sensitive information ratio we propose to do the Cauchy-Schwarz step differently and as follows (using the fact that $\ell_t(i)^2 \leq \ell_t(i)$):

$$\begin{aligned} r_t = \mathbb{E}_t[\langle \ell_t, p_t - p_{t+1} \rangle] &\leq \sqrt{\mathbb{E}_t \left[\sum_{i=1}^d \ell_t(i) p_t(i) \right] \cdot \mathbb{E}_t \left[\sum_{i=1}^d \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \right]} \\ &= \sqrt{\mathbb{E}_t[\langle \ell_t, p_t \rangle] \cdot \mathbb{E}_t[\chi^2(p_t, p_{t+1})]}\,, \end{aligned} \tag{5}$$

where $\chi^2(p,q) = \sum_{i=1}^d \frac{(p(i)-q(i))^2}{p(i)}$ is the chi-squared divergence. Thus, to control the scale-sensitive information ratio (4), it only remains to relate the chi-squared divergence to the relative entropy. Unfortunately it is well-known that in general one only has $\mathrm{Ent}(q,p) \leq \chi^2(p,q)$ (which is the opposite of the inequality we need). Somewhat surprisingly we show that the reverse inequality in fact holds up to a factor of two true for a slightly weaker form of the chi-squared divergence, which turns out to be sufficient for our needs:

Lemma 8 *For $p, q \in \mathbb{R}_+^d$ define the* positive chi-squared divergence χ_+^2 *by*

$$\chi_+^2(p,q) = \sum_{i : p(i) \geq q(i)} \frac{(p(i) - q(i))^2}{p(i)}\,.$$

Also we denote $\mathrm{Ent}(p,q) = \sum_{i=1}^d \left(p(i) \log(p(i)/q(i)) - p(i) + q(i) \right)$. *Then one has*

$$\chi_+^2(p,q) \leq 2\mathrm{Ent}(q,p)\,.$$

Proof Consider the function $f_t(s) = s\log(s/t) - s + t$, and observe that $f_t''(s) = 1/s$. In particular f_t is convex, and for $s \leq t$ it is $\frac{1}{t}$-strongly convex. Moreover one has $f_t'(t) = 0$. This directly implies:

$$f_t(s) \geq \frac{1}{2t}(t-s)_+^2\,,$$

which concludes the proof. ■

Combining Lemma 8 with (5) (where we replace r_t by r_t^+) one obtains the following:

Lemma 9 *In the expert setting, Thompson Samping's scale-sensitive information ratio (4) satisfies $\Lambda_t \leq 2$ for all t.*

Using Lemma 9 in Proposition 7 we arrive at the following new regret bound for Thompson Sampling:

Theorem 10 *In the expert setting Thompson Sampling satisfies for any prior distribution:*

$$\mathbb{E}[R_T] \leq \sqrt{2\mathbb{E}[L^*] \cdot H(p_1)} + 2H(p_1)\,.$$

3. Combinatorial setting and coordinate entropy

We now return to the general combinatorial setting, where the action set $\mathcal{A}$ is a subset of $\{A \in \{0,1\}^d : \|A\|_1 = m\}$, and we continue to focus on the full information game. Recall that, as described in Theorem 3, Russo and Van Roy's analysis yields in this case the suboptimal regret bound $\widetilde{O}(m^{3/2}\sqrt{T})$ (the optimal bound is $m\sqrt{T}$). We first argue that this suboptimal bound comes from basing the analysis on the standard Shannon entropy. We then propose a different analysis based on the *coordinate entropy*.

3.1. Inadequacy of the Shannon entropy

Let us consider the simple scenario where $\mathcal{A}$ is the set of indicator vectors for the sets $a_k = \{1 + (k-1) \cdot m, \ldots, k \cdot m\}$, $k \in [d/m]$. In other words, the action set consists of $\frac{d}{m}$ disjoint intervals of size m. This problem is equivalent to a classical expert setting with d/m actions, and losses with values in $[0, m]$. In particular there exists a prior distribution such that any algorithm must suffer regret $m\sqrt{T \log(d/m)} \geq m\sqrt{TH(p_1)}$ (the lower bound comes from the fact that there is only d/m available actions).

Thus we see that, unless the regret bound reflects some of the structure of the action set $\mathcal{A} \subset \{0,1\}^d$ (besides the fact that elements have m non-zero coordinates), one cannot hope for a better regret than $m\sqrt{TH(p_1)}$. For larger action sets, $H(p_1)$ could be as large as $m \log(d/m)$. Thus, if we are to obtain a (m, T) dependent regret bound via the entropy of the optimal action set, the best possible bound will be $m^{3/2}\sqrt{T}$. However the optimal rate for this online learning problem is known to be $\widetilde{O}(m\sqrt{T})$. This suggests that the Shannon entropy is not the right measure of uncertainty in this combinatorial setting, at least if we expect Thompson Sampling to perform optimally.

Interestingly a similar observation was made in (ABL14) where it was shown that the regret for the standard multiplicative weights algorithm is also lower bounded by the suboptimal rate $m^{3/2}\sqrt{T}$. The connection to the present situation is that standard multiplicative weights corresponds to mirror descent with the Shannon entropy. To obtain an optimal algorithm, (KWK10; ABL14) proposed to use mirror descent with a certain *coordinate entropy*. We show next that basing the analysis of Thompson Sampling on this coordinate entropy allows us to prove optimal guarantees.

3.2. Coordinate entropy analysis

For any vector $v = (v_1, v_2, \ldots, v_d) \in [0,1]^d$, we define its *coordinate entropy* $H^c(v)$ to simply be the sum of the entropies of the individual coordinates:

$$H^c(v) = \sum_{i=1}^{d} H(v_i) = -\sum_{i=1}^{d} v_i \log(v_i) + (1 - v_i) \log(1 - v_i).$$

For a $\{0,1\}^d$-valued random variable such as a^*, we define $H^c(a^*) = H^c(\mathbb{E}[a^*])$. Equivalently, the coordinate entropy $H^c(a^*)$ is the sum of the (ordinary) entropies of the d Bernoulli random variables $1_{i \in a^*}$.

This definition allows us to consider the information gain in each event $[i \in a^*]$ separately in the information-theoretic analysis via $I_t^c = H_t^c(p_t) - \mathbb{E}_t[H_t^c(p_{t+1})]$, denoting now $p_t = \mathbb{E}_t[a_t]$. By inspecting our earlier proof one easily obtains in the combinatorial setting

$$(r_t^+)^2 \leq 2\langle p_t, \ell_t \rangle \cdot \mathbb{E}_t[\mathrm{Ent}(p_{t+1}, p_t)] = 2\langle p_t, \ell_t \rangle \cdot \mathbb{E}_t[I_t^c] \,. \tag{6}$$

As a result, the scale-sensitive information ratio with coordinate entropy is $\Lambda_t^c := \frac{(r_t^+)^2)}{I_t^c \cdot \mathbb{E}_t[\langle \ell_t, a_t \rangle]} \leq 2$. Therefore

$$\mathbb{E}[R_T] \leq \sqrt{2\mathbb{E}[L^*] H^c(p_1)} + 2H(p_1).$$

To establish the first half of Theorem 4 we just need to estimate $H(p_1)$. By Jensen's inequality, we have

$$H^c(p_1) \leq H^c\left(\frac{m}{d}, \frac{m}{d}, \ldots, \frac{m}{d}\right) = m \log\left(\frac{d}{m}\right) + (d - m) \log\left(\frac{d}{d-m}\right).$$

Using the inequality $\log(1 + x) \leq x$ on the second term we obtain

$$H^c(p_1) \le m \log\left(\frac{d}{m}\right) + m$$

This gives the claimed estimate

$$\mathbb{E}[R_T] = O(\sqrt{m \log(d/m)\mathbb{E}[L^*]}).$$

Remark 11

The fact we use the coordinate entropy suggests that it is unnecessary to leverage information from correlations between different arms, and we can essentially treat them as independent. In fact, our proofs for Thompson Sampling apply to any algorithm which observes arm i at time t with probability $p_t(i \in a^*)$. This remark extends to the thresholded variants of Thompson Sampling we discuss at the end of the paper.

4. Bandit

Now we return to the $m = 1$ setting and consider the case of bandit feedback. We again begin by recalling the analysis of Russo and Van Roy, and then adapt it in analogy with the scale-sensitive framework. For most of this section, we require that an almost sure upper bound $L^* \le \underline{L}^*$ for the loss of the best action is given to the player. Under this assumption we show that Thompson Sampling obtains a regret bound $\widetilde{O}(\sqrt{H(p_1)d\underline{L}^*})$, by using a bandit analog of the method in the previous section. This estimate can be improved with the method of (ZL19) which shows how to analyze Thompson Sampling based on online stochastic mirror descent. By using a logarithmic regularizer, we obtain a regret bound depending only on $\mathbb{E}[L^*]$, i.e. *without* the assumption $L^* \le \underline{L}^*$, matching the statement of Theorem 4.

4.1. The Russo and Van Roy Analysis for Bandit Feedback

In the bandit setting we cannot bound the regret by the movement of p_t. Indeed, the calculation (3) relies on the fact that ℓ_t is known at time $t + 1$ which is only true for full feedback. However, a different information theoretic calculation gives a good estimate.

Lemma 12 ((RVR16))

In the bandit setting, Thompson Sampling's information ratio satisfies $\Gamma_t \le d$ for all t. Therefore it has expected regret $\mathbb{E}[R_T] \le \sqrt{dTH(p_1)}$.

Proof We set $\bar{\ell}_t(i) = \mathbb{E}_t[\ell_t(i)]$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t[\ell_t(i)|a^* = j]$. Then we have the calculation

$$r_t = p_t \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \le \sqrt{d \cdot \left(\sum_i p_t(i)^2 \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2\right)}$$

$$\le \sqrt{d \cdot \left(\sum_i p_t(i)^2 \text{Ent}[\bar{\ell}_t(i,i), \bar{\ell}_t(i)]\right)}.$$

By Lemma 13 below, this means

$$r_t \le \sqrt{d \cdot I_t}$$

which is equivalent to $\Gamma_t \le d$. ∎

The following lemma is a generalization of a calculation in (RVR16). We leave the proof to the Appendix.

Lemma 13

Suppose a Bayesian player is playing a semi-bandit game with a hidden subset S of arms. Each round t, the player picks some subset a_t of arms and observes all their losses. Define $p_t(i \in S) = \mathbb{P}[i \in S]$, and $\hat{p}_t(i) = \mathbb{P}[i \in a_t]$. Let $\bar{\ell}_t(i, i \in S) = \mathbb{E}[\ell_t(i) | i \in S]$. Then with $I_t^c(\cdot)$ the coordinate information gain we have

$$\sum_i \hat{p}_t(i) p_t(i \in S) \mathrm{Ent}[\bar{\ell}_t(i, i \in S), \bar{\ell}_t(i)] \leq I_t^c[S].$$

The next lemma is a scale-sensitive analog of an information ratio bound for partial feedback. However, getting from such a statement to a regret bound is a bit more involved in our small loss setting so we do not try to push the analogy too far.

Lemma 14

In the setting of Lemma 13, we have:

$$\sum_i \hat{p}_t(i) p_t(i \in S) \left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i, i \in S))_+^2}{\bar{\ell}_t(i)} \right) \leq 2 \cdot I_t^c[S].$$

4.2. General Theorem on Perfectly Bayesian Agents

Here we state a theorem on the behavior of a Bayesian agent in an online learning environment. In the next subsection we use it to give a nearly optimal regret bound for Thompson Sampling with bandit feedback. This theorem is stated in a rather general way in order to encompass the semi-bandit case as well as the thresholded version of Thompson Sampling discussed later. The proof goes by controlling the errors of unbiased and negatively biased estimators for the losses using a concentration inequality. Then we argue that because these estimators are accurate with high probability, a Bayesian agent will usually believe them to be accurate, even though this accuracy circularly depends on the agent's past behavior. We relegate the detailed proof to the Appendix.

Theorem 15

Consider an online learning game with arm set $[d]$ and random set of losses $\ell_t(i)$. Suppose also that the player is given the distribution from which the loss sequence is sampled, i.e. an accurate prior distribution. Assume there always exists an action with total loss at most $\underline{L}^*$. Each round, the player plays some subset a_t of actions, and pays/observes the loss for each of them. Let $p_t(i) = \mathbb{P}_t[i \in a^*]$ be the time-t probability that i is one of the optimal arms and $\hat{p}_t(i) = \mathbb{P}_t[i \in a_t]$ the probability that the player plays arm i in round t. We suppose that there exist constants $\frac{1}{\underline{L}^*} \leq \gamma_1 \leq \gamma_2$ and a time-varying partition $[d] = \mathcal{R}_t \cup \mathcal{C}_t$ of the action set into rare and common arms such that:

1. *If $i \in \mathcal{C}_t$, then $\hat{p}_t(i), p_t(i) \geq \gamma_1$.*
2. *If $i \in \mathcal{R}_t$, then $\hat{p}_t(i) \leq p_t(i) \leq \gamma_2$.*

Then the following statements hold for every i.

A) *The expected loss incurred by the player from rare arms is at most*

$$\mathbb{E}\left[\sum_{t:\, i\in\mathcal{R}_t} \hat{p}_t(i)\ell_t(i)\right] \leq 2\gamma_2 \underline{L}^* + 8\log(T) + 4.$$

B) *The expected total loss that arm i incurs while it is common is at most*

$$\mathbb{E}\left[\sum_{t:\, i\in\mathcal{C}_t} \ell_t(i)\right] \leq \underline{L}^* + 2\left(\log\left(\frac{1}{\gamma_1}\right) + 10\right)\sqrt{\frac{\underline{L}^*}{\gamma_1}}$$

4.3. First-Order Regret for Bandit Feedback

As Theorem 15 alluded to, we split the action set into *rare* and *common* arms for each round. Rare arms are those with $p_t(i) \leq \gamma$ for some constant $\gamma > 0$, while common arms have $p_t(i) > \gamma$. Note that an arm can certainly switch from rare to common and back over time. We correspondingly split the loss function into

$$\ell_t(i) = \ell_t^{\mathcal{R}}(i) + \ell_t^{\mathcal{C}}(i)$$

via $\ell_t^{\mathcal{R}}(i) = \ell_t(i) 1_{p_t(i)\leq\gamma}$ and $\ell_t^{\mathcal{C}}(i) = \ell_t(i) 1_{p_t(i)>\gamma}$.

Now we are ready to prove the first-order regret bound for bandits.

Theorem 16

Suppose that the best expert almost surely has total loss at most $\underline{L}^*$. Then Thompson Sampling with bandit feedback obeys the regret estimate

$$\mathbb{E}[R_T] \leq O\left(\sqrt{H(p_1)d\underline{L}^*} + d\log^2(\underline{L}^*) + d\log(T)\right).$$

Proof Fix $\gamma > 0$ and define $\mathcal{R}_t$ and $\mathcal{C}_t$ correspondingly. We split off the rare arm losses at the start of the analysis:

$$\mathbb{E}[R_T] \leq \mathbb{E}\left[\sum_t r_t^+\right] = \mathbb{E}\left[\sum_t p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))_+\right]$$
$$\leq \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] + \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))_+\right].$$

The first term is bounded by Theorem 15A with the rare/common partition above, $\gamma_1 = \gamma_2 = \gamma$, and $\hat{p}_t(i) = p_t(i)$. For the second term, again using Cauchy-Schwarz and then Lemmas 8 and 14 gives:

$$\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))_+\right] \leq \sqrt{\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} \bar{\ell}_t(i)\right]\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)^2\cdot\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))_+^2}{\bar{\ell}_t(i)}\right]}$$
$$\leq \sqrt{2\cdot\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} \ell_t(i)\right]\cdot H(p_1)}.$$

Substituting in the conclusion of Theorem 15B gives:

$$\mathbb{E}[R_T] \leq d(2\gamma \underline{L}^* + 8\log(T) + 4) + \sqrt{H(p_1)\left(\underline{L}^* + 2\left(\log\left(\frac{1}{\gamma}\right) + 10\right)\sqrt{\frac{\underline{L}^*}{\gamma}}\right)}.$$

Taking $\gamma = \frac{\log^2(\underline{L}^*)}{\underline{L}^*}$ gives the desired estimate.

■

4.4. Improved Estimates Beyond Shannon Entropy

In recent work (ZL19), it is shown that Thompson sampling can be analyzed using any mirror map, with the same guarantees as online stochastic mirror descent. (See also (LS19) which essentially improves the Russo and Van Roy entropic bound using Tsallis entropy.) Their work is compatible with our methods for first order analysis, allowing for further refinements. By using the Tsallis entropy we remove the $H(p_1)$ dependence in Theorem 16, and also gain the potential for polynomial-in-d savings for an informative prior. By using the log barrier we obtain a small loss bound depending only on $\mathbb{E}[L^*]$ instead of requiring an almost sure upper bound $\underline{L}^*$. The proofs follow a similar structure to that of Theorem 16 and we leave them to the appendix.

Definition 17

For $\alpha \in (0,1)$ we define the α-Tsallis entropy of a probability vector p to be

$$H_\alpha(p) = \frac{(\sum_i p_i^\alpha) - 1}{\alpha(1-\alpha)}.$$

Note that with d actions, $H_\alpha(p) \leq \frac{d^{1-\alpha}}{\alpha(1-\alpha)}$.

Theorem 18

Suppose that the best expert almost surely has total loss at most $\underline{L}^*$. Then Thompson Sampling with bandit feedback obeys the regret estimate

$$\mathbb{E}[R_T] \leq O\left(\sqrt{H_\alpha(p_1)d^\alpha \underline{L}^*} + H_\alpha(p_1)d^\alpha + d\log^2(\underline{L}^*) + d\log(T)\right).$$

For any fixed α, this gives a worst-case over p_1 upper bound of

$$\mathbb{E}[R_T] \leq O_\alpha\left(\sqrt{d\underline{L}^*} + d\log^2(\underline{L}^*) + d\log(T)\right).$$

Theorem 19

Thompson Sampling with bandit feedback obeys the regret estimate

$$\mathbb{E}[R_T] = O(\sqrt{d\mathbb{E}[L^*]\log(T)} + d\log(T)$$

We observe that for a highly informative prior, Theorem 18 may be much tighter than a worst case bound. For example if $p_1(i) \lesssim i^{-\beta}$ for some $\beta > 1$, then for $\alpha \geq \frac{1}{\beta}$ we will have $H_\alpha(p_1)$ bounded independently of d. Hence the main term of the regret will be $O_\alpha(\sqrt{d^\alpha \underline{L}^*})$, meaning we save a multiplicative power of d!

We also note that Theorem 19 actually does not require Theorem 15 so its proof is somewhat simpler than Theorem 18 given the connection to mirror descent. However the $\underline{L}^*$ dependent results have the interesting advantage of leading to fully T-independent regret with thresholded Thompson Sampling in the next section.

5. Semi-bandit, Thresholding, and Graph Feedback

In this section, we consider three extensions of the previous results. We first combine the combinatorial setting with bandit feedback, the so-called semi-bandit model. Next, we show how to obtain T-independent regret with a thresholded version of Thompson Sampling when there is an almost sure upper bound $L^* \leq \underline{L}^*$. Finally we show that coordinate entropy allows for $\sqrt{T}$ type regret bounds for Thompson sampling under graph feedback.

5.1. Semi-bandit

We now consider semi-bandit feedback in the combinatorial setting, combining the intricacies of the previous two sections. We again have an action set $\mathcal{A}$ contained in the set $\{a \in \{0,1\}^d : ||a||_1 = m\}$, but now we observe the m losses of the arms we played. A natural generalization of the bandit $m = 1$ proof to higher m yields a first-order regret bound of $\widetilde{O}(\sqrt{md\underline{L}^*})$. However, we give a refined analysis using an additional trick of ranking the m arms in a^* by their total loss and performing an information theoretic analysis on a certain set partition of these m optimal arms. This method allows us to obtain a $\widetilde{O}(\sqrt{d\underline{L}^*})$ regret bound for the semi-bandit regret. The analyses based on other mirror maps extend as well. We leave the proofs to the Appendix.

Theorem 20 *Suppose that the best combinatorial action almost surely has total loss at most $\underline{L}^*$. Then Thompson sampling with semi-bandit feedback obeys the regret estimate*

$$\mathbb{E}[R_T] \leq O\left(\log(m)\sqrt{d\underline{L}^*} + md^2\log^2(\underline{L}^*) + d\log(T)\right).$$

The upper bound with the log barrier is identical to the semibandit case.

Theorem 21 *Thompson sampling with semi-bandit feedback obeys the regret estimate*

$$\mathbb{E}[R_T] \leq O\left(\sqrt{d\mathbb{E}[L^*]\log(T)} + d\log(T)\right).$$

5.2. Thresholded Thompson Sampling

Unlike in the full-feedback case, our first-order regret bound for bandit Thompson Sampling has an additive $O(d\log(T))$ term. Thus, even when an upper bound $L^* \leq \underline{L}^*$ is known, the regret is T-independent. In fact, some mild T-dependence is inherent - an example is given in the Appendix.

However, this mild T-dependence can be avoided by using *Thresholded Thompson Sampling*. In Thresholded Thompson Sampling, the rare arms are *never* played, and the probabilities for the other arms are scaled up correspondingly. More precisely, for $\gamma < \frac{1}{d}$, the γ-thresholded Thompson Sampling is defined by letting $\mathcal{R}_t = \{i : p_t(i) \leq \gamma\}$ and playing at time t from the distribution

$$\hat{p}_t(i) = \begin{cases} 0 & \text{if } i \in \mathcal{R}_t \\ \frac{p_t(i)}{1-\sum_{j\in\mathcal{R}_t} p_t(j)} & \text{if } i \in \mathcal{C}_t. \end{cases}$$

This algorithm parallels the work (LST18) which uses an analogous modification of the EXP3 algorithm to obtain a first-order regret bound. Thresholded semi-bandit Thompson Sampling is defined similarly, where we only allow action sets containing no rare arms.

Theorem 22

Suppose that the best action almost surely has total loss at most $\underline{L}^*$. Thompson Sampling for bandit feedback, thresholded with $\gamma = \frac{\log^2(\underline{L}^*)}{\underline{L}^*} \leq \frac{1}{2d}$, has expected regret

$$\mathbb{E}[R_T] = O\left(\sqrt{d\underline{L}^*} + d\log^2(\underline{L}^*)\right)$$

Theorem 23

Suppose that the best combinatorial action almost surely has total loss at most $\underline{L}^*$. Thompson Sampling for semi-bandit feedback, thresholded with $\gamma = \frac{m\log^2(\underline{L}^*)}{\underline{L}^*} \leq \frac{1}{2d}$, has expected regret

$$\mathbb{E}[R_T] = O\left(\log(m)\sqrt{d\underline{L}^*} + md\log^2(\underline{L}^*)\right).$$

5.3. Graphical Feedback

Here we consider the problem of online learning with graphical feedback. This model interpolates between full-feedback and bandits by embedding the actions as vertices of a (possibly directed) graph G with d vertices. If there is an edge $i \to j$, then playing action i allows one to observe action j. We assume that all vertices have self-loops, i.e. that we always observe the loss we pay. Without this assumption, the optimal regret might be $\widetilde{\Theta}(T^{2/3})$ even if every vertex is observable, see (ACBDK15).

Previous work such as (LZS18; TDD17) analyzed the performance of Thompson Sampling for these tasks, giving $O(\sqrt{T})$-type regret bounds which scale with certain statistics of the graph. However, their analyses only applied for stochastic losses rather than adversarial losses. In this section, we outline why their analysis applies to the adversarial case as well.

5.3.1. Entropy for Graph Feedback

Here we generalize the analysis of (LZS18) to the adversarial setting. As above, let G be a possibly directed feedback graph on d vertices, with $\alpha = \alpha(G)$ its independence number (the size of the maximum independent set). A key point is:

Lemma 24 ((MS11), Lemma 3)

For any probability distribution π on $V(G)$ we have (under the convention $0/0 = 0$):

$$\sum_i \frac{\pi(i)}{\sum_{j\in\{i\}\cup N(i)} \pi(j)} \leq \alpha.$$

From this, following (LZS18) we obtain:

Proposition 25

The coordinate information ratio of Thompson Sampling on an undirected graph G is at most $\alpha(G)$.

Proof

Let $p_t(i)$ be as usual for a vertex i and $q_t(i) = \sum_{j\in\{i\}\cup N(i)} p_t(i)$ the probability to observe $\ell_t(i)$. Then:

$$\alpha \cdot I_t^c \geq \left(\sum_i \frac{p_t(i)}{q_t(i)}\right)\left(\sum_i p(i)q(i)(\ell_t(i) - \ell_t(i,i))^2\right) \geq R^2.$$

■

In the case of a directed graph, the natural analog of $\alpha(G)$ is the maximum value of

$$\sum_i \frac{\pi(i)}{\sum_{j\in\{i\}\cup N^{in}(i)} \pi(j)}$$

which is equal to $mas(G)$, the size of the maximal acyclic subgraph of G. However, as noted in (LZS18), if we assume $\pi_t(i) \geq \varepsilon$ then (ACBDK15) gives the upper bound

$$\sum_i \frac{\pi(i)}{\sum_{j\in\{i\}\cup N^{in}(i)} \pi(j)} \leq 4\left(\alpha \cdot \log\left(\frac{4d}{\alpha\varepsilon}\right)\right).$$

By using the fact that $\varepsilon = o(T^{-2})$ additional exploration has essentially no effect on the expected regret (as it induces $o(T)$ total variation distance betwen the two algorithms), we obtain a α-dependent bound for directed graphs as well:

Theorem 26

Thompson Sampling on a sequence G_t of undirected graphs achieves expected regret

$$\mathbb{E}[R_T] = O\left(\sqrt{H^c(p_1)\sum_{t=1}^{T}\alpha(G_t)}\right).$$

Moreover Thompson Sampling on a sequence G_t of directed graphs achieves expected regret

$$\mathbb{E}[R_T] = O\left(\sqrt{H^c(p_1)\log(dT)\sum_{t=1}^{T}\alpha(G_t)}\right).$$

As in (LZS18), this analysis applies even when the Thompson sampling algorithm does not know the graphs G_t, but only observes the relevant neighborhood feedback after choosing each action a_t.

6. Negative Results for Thompson Sampling

In this section we give a few negative results. Theorem 27 states that Thompson Sampling against an arbitrary prior may have $\Omega(T)$ regret a constant fraction of the time (but will therefore also have $-\Omega(T)$ regret a constant fraction of the time). By contrast, there exist algorithms which have low regret with high probability even in the frequentist setting (ACBFS02). Bridging this gap with a variant of Thompson Sampling would be very interesting.

Theorem 28 shows that the slight T dependence in our Thompson Sampling guarantees was necessary. Recall that even in Theorem 18 there was an additive $d\log(T)$ term in the expected regret. The lower bound in Theorem 28 seems correspond to an additive $d\log^*(T)$ term, where $\log^*$ is the inverse to the tower function.

Theorem 27

There exist prior distributions for which Thompson Sampling achieves $\Omega(T)$ regret a constant fraction of the time, with either full or bandit feedback.

Theorem 28

There exist prior distributions against which Thompson Sampling achieves $\Omega(d\underline{L}^*)$ expected regret for very large T with bandit feedback, even given the value $\underline{L}^*$.

6.1. Lower Bound for Contextual Bandit

Recall that contextual bandit is equivalent to graph feedback where:

- The graphs change from round to round.
- All graphs are vertex-disjoint unions of at most K cliques.
- The losses for a round are constant within cliques.

The existence of an algorithm achieving $O(\sqrt{L^*})$ regret for contextual bandits was posed in (AKL+17) and resolved positively in (AZBL18). However the algorithm given is highly computationally intractable. Here we show that Thompson Sampling does not match this guarantee.

Theorem 29

It is possible that Thompson Sampling achieves, with high probability, loss $\Omega(\sqrt{T})$ for a contextual bandit problem with $L^* = 0$ optimal loss, $K = 2$ cliques, and $d = O(\sqrt{T})$ total arms.

References

[AAGO06] C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory (ALT)*, 2006.

[ABL14] J.Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39:31–45, 2014.

[ACBDK15] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Annual Conference on Learning Theory*, volume 40. Microtome Publishing, 2015.

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[AG12] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 23, 2012.

[AKL+17] Alekh Agarwal, Akshay Krishnamurthy, John Langford, Haipeng Luo, et al. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, pages 4–7, 2017.

[AZBL18] Zeyuan Allen-Zhu, Sebastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, pages 186–194, 2018.

[BDKP15] S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, 2015.

[CBFH$^+$97] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

[CL11] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[Fre75] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

[KWK10] W. Koolen, M. Warmuth, and J. Kivinen. Hedging structured concepts. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

[LS19] Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 2111–2139, 2019.

[LST18] T. Lykouris, K. Sridharan, and E. Tardos. Small-loss bounds for online learning with partial information. In *Proceedings of the 31st Annual Conference on Learning Theory (COLT)*, 2018.

[LTW20] Thodoris Lykouris, Eva Tardos, and Drishti Wali. Feedback graph regret bounds for thompson sampling and ucb. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT)*, 2020.

[LZS18] Fang Liu, Zizhan Zheng, and Ness Shroff. Analysis of thompson sampling for graphical bandits without the graphs. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[MS11] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

[RVR16] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[TDD17] Aristide CY Tossou, Christos Dimitrakakis, and Devdatt P Dubhashi. Thompson sampling for stochastic bandits with graph feedback. In *AAAI*, pages 2660–2666, 2017.

[Tho33] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, 25:285–294, 1933.

[ZL19] Julian Zimmert and Tor Lattimore. Connections between mirror descent, thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems 32 (NIPS)*, 2019.

Appendix A.

A.1. Proof of Lemmas 13 and 14

Proof of Lemma 13:

We first claim that the relative entropy

$$\mathrm{Ent}[\bar{\ell}_t(i, i \in S), \bar{\ell}_t(i)]$$

is at most the entropy decrease in the law $\mathcal{L}_t(\ell_t(i))$ of the random variable $\ell_t(i)$ from being given that $i \in S$. Indeed, let $\widetilde{\ell}_t(i)$ be a $\{0,1\}$-valued random variable with expected value $\bar{\ell}_t(i)$ and conditionally independent of everything else. By definition,

$$\mathrm{Ent}[\bar{\ell}_t(i, i \in S), \bar{\ell}_t(i)]$$

is exactly the information gain in $\mathcal{L}_t(\widetilde{\ell}_t(i))$ upon being told that $j \in S$. Since $\widetilde{\ell}_t(i)$ is a noisy realization of $\ell_t(i)$, the data processing inequality implies that the information gain of $\mathcal{L}_t(\ell_t(i))$ is more than the information gain in $\mathcal{L}_t(\widetilde{\ell}_t(i))$ which proves the claim.

Now, continuing, we have that

$$p_t(i \in S)\mathrm{Ent}[\bar{\ell}_t(i, i \in S), \bar{\ell}_t(i)]$$

is at most the entropy decrease in $\ell_t(i)$ from being given whether or not $i \in S$. Therefore

$$\hat{p}_t(i)p_t(i \in S)\mathrm{Ent}[\bar{\ell}_t(i, i \in S), \bar{\ell}_t(i)] \le I[\ell_t(i)1_{i \in a_t}, 1_{i \in S}]$$

$$\le I[(a_t, \boldsymbol{\ell}_t(a_t)), 1_{i \in S}] = I_t[1_{i \in S}].$$

■

Proof of Lemma 14:

By Lemma 8 we have:

$$\sum_i \hat{p}_t(i)p_t(i \in S)\left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i, i \in S))_+^2}{\bar{\ell}_t(i)}\right) \le 2\sum_i \hat{p}_t(i)p_t(i)\big(Ent(\bar{\ell}_t(i, i \in S), \bar{\ell}_t(i)) - \bar{\ell}_t(i, i \in S) + \bar{\ell}_t(i)\big)$$

and

$$\sum_i \hat{p}_t(i)p_t(i \notin S)\left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i, i \notin S))_+^2}{\bar{\ell}_t(i)}\right) \le 2\sum_i \hat{p}_t(i)p_t(i \notin S)\big(Ent(\bar{\ell}_t(i, i \notin S), \bar{\ell}_t(i)) - \bar{\ell}_t(i, i \notin S) + \bar{\ell}_t(i)\big)$$

Summing and noting that

$$p_t(i \in S)\bar{\ell}_t(i, i \in S) + p_t(i \notin S)\bar{\ell}_t(i, i \notin S) = p_t(i \in S)\bar{\ell}_t(i) + p_t(i \notin S)\bar{\ell}_t(i) = \bar{\ell}_t(i)$$

we obtain

$$\sum_i \hat{p}_t(i)p_t(i \in S)\left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i, i \in S))_+^2}{\bar{\ell}_t(i)}\right) + \sum_i \hat{p}_t(i)p_t(i \notin S)\left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i, i \notin S))_+^2}{\bar{\ell}_t(i)}\right)$$

$$\le 2\sum_i \hat{p}_t(i)\Big(p_t(i\in S)Ent(\bar{\ell}_t(i,i\in S),\bar{\ell}_t(i)) + p_t(i\notin S)Ent(\bar{\ell}_t(i,i\notin S),\bar{\ell}_t(i))\Big).$$

Since $\hat{p}_t(i)$ is the chance to observe $\ell_t(i)$, and it is multiplied by a lower bound for the information gain on the event $\{i\in S\}$ from observing $\ell_t(i)$, the RHS is bounded by $2\cdot I_t^c[S]$. The extra term we added is non-negative so we conclude the lemma.

■

A.2. Proof of Theorem 15

Here we prove Theorem 15. Recall the statement:

Theorem 15

Consider an online learning game with arm set $[d]$ and random set of losses $\ell_t(i)$. Suppose also that the player is given the distribution from which the loss sequence is sampled, i.e. an accurate prior distribution. Assume there always exists an action with total loss at most $\underline{L}^*$. Each round, the player plays some subset a_t of actions, and pays/observes the loss for each of them. Let $p_t(i)=\mathbb{P}_t[i\in a^*]$ be the time-t probability that i is one of the optimal arms and $\hat{p}_t(i)=\mathbb{P}_t[i\in a_t]$ the probability that the player plays arm i in round t. We suppose that there exist constants $\frac{1}{\underline{L}^*}\le\gamma_1\le\gamma_2$ and a time-varying partition $[d]=\mathcal{R}_t\cup\mathcal{C}_t$ of the action set into rare and common arms such that:

1. *If $i\in\mathcal{C}_t$, then $\hat{p}_t(i),p_t(i)\ge\gamma_1$.*
2. *If $i\in\mathcal{R}_t$, then $\hat{p}_t(i)\le p_t(i)\le\gamma_2$.*

Then the following statements hold for every i.

A*) The expected loss incurred by the player from rare arms is at most*

$$\mathbb{E}\left[\sum_{t:\, i\in\mathcal{R}_t}\hat{p}_t(i)\ell_t(i)\right]\le 2\gamma_2\underline{L}^*+8\log(T)+4.$$

B*) The expected total loss that arm i incurs while it is common is at most*

$$\mathbb{E}\left[\sum_{t:\, i\in\mathcal{C}_t}\ell_t(i)\right]\le\underline{L}^*+2\left(\log\left(\frac{1}{\gamma_1}\right)+10\right)\sqrt{\frac{\underline{L}^*}{\gamma_1}}$$

The following notations will be relevant to our analysis. Some have been defined in the main body, while some are only used in the Appendix.

$\ell_t^{\mathcal{R}}(i)=\ell_t(i)\cdot 1_{i\in\mathcal{R}_t}$	$u_t^{\mathcal{R}}(i)=\frac{\ell_t^{\mathcal{R}}(i)\cdot 1_{i_t=i}}{\gamma_2}$	$L_t^{\mathcal{R}}(i)=\sum_{s\le t}\ell_s^{\mathcal{R}}(i)$	$U_t^{\mathcal{R}}(i)=\sum_{s\le t}u_s^{\mathcal{R}}(i)$
$\ell_t^{\mathcal{C}}(i)=\ell_t(i)\cdot 1_{i\in\mathcal{C}_t}$	$u_t^{\mathcal{C}}(i)=\frac{\ell_t^{\mathcal{C}}(i)\cdot 1_{i_t=i}}{\hat{p}_t(i)}$	$L_t^{\mathcal{C}}(i)=\sum_{s\le t}\ell_s^{\mathcal{C}}(i)$	$U_t^{\mathcal{C}}(i)=\sum_{s\le t}u_s^{\mathcal{C}}(i)$

The ℓ_t variables are the instantaneous rare/common losses of an arm, while the L_t variables track the total loss. The u_t variables are underbiased/unbiased estimates of the ℓ_t while the U_t variables are underbiased/unbiased estimates of the L_t.

To control the error of the estimators U_t we rely on Freedman's inequality ((Fre75)), a refinement of Hoeffding-Azuma which is more efficient for highly assymmetric summands.

Theorem 30 (Freedman's Inequality)

Let $S_t = \sum_{s \le t} x_s$ be a martingale sequence, so that

$$\mathbb{E}[x_s | \mathcal{F}_{s-1}] = 0.$$

Suppose that we have a uniform estimate $x_s \le M$. Also define the conditional variance

$$W_s = Var[X_s | \mathcal{F}_{s-1}]$$

and set $V_t = \sum_{s \le t} W_s$ to be the total variance accumulated so far.

Then with probability at least $1 - e^{-\frac{a^2}{2b+Ma}}$, we have $S_t \le a$ for all t with $V_t \le b$.

The following extension to supermartingales is immediate by taking the Doob-Meyer decomposition of a supermartingale as a martingale plus a decreasing predictable process.

Corollary 31 *Let $S_t = \sum_{s \le t} x_s$ be a supermartingale sequence, so that $\mathbb{E}[x_s | \mathcal{F}_{s-1}] \le 0$. Suppose that we have a uniform estimate $x_s - \mathbb{E}[x_s | \mathcal{F}_{s-1}] \le M$. Also define the conditional variance*

$$W_s = Var[X_s | \mathcal{F}_{s-1}]$$

and set $V_t = \sum_{s \le t} W_s$ to be the total variance accumulated so far.

Then with probability at least $1 - e^{-\frac{a^2}{2b+Ma}}$, we have $S_t \le a$ for all t with $V_t \le b$.

Towards proving the two claims in Theorem 15 we first prove two lemmas. They follow directly from proper applications of Freedman's Theorem or its corollary.

Lemma 32 *In the context of Theorem 15, with probability at least $1 - \frac{2}{T^2}$, for all t with $L_t^{\mathcal{R}}(i) \le \underline{L}^*$ we have*

$$U_t^{\mathcal{R}}(i) \le 2\underline{L}^* + \frac{8 \log T}{\gamma_2}.$$

Lemma 33

In the context of Theorem 15, fix constants $\lambda \ge 2$ and $\widetilde{L} > 0$ and assume $\gamma_1 \ge \frac{1}{\widetilde{L}}$. With probability at least $1 - 2e^{-\lambda/2}$, for all t with $L_t^{\mathcal{C}}(i) \le \widetilde{L}$ we have

$$U_t^{\mathcal{C}}(i) \le L_t^{\mathcal{C}}(i) + \lambda \sqrt{\frac{\widetilde{L}}{\gamma_1}}.$$

Remark 34

This second lemma has no dependence on $\underline{L}^*$ and holds with $\underline{L}^* = \infty$. For proving Theorem 15 we will simply take $\widetilde{L} = \underline{L}^*$. We will need to apply this lemma with $\widetilde{L} \ne \underline{L}^*$ for the semi-bandit analog.

Proof of Lemma 32:

We analyze the (one-sided) error in the underestimate $U_t^{\mathcal{R}}(i)$ for $L_t^{\mathcal{R}}(i)$. Define the supermartingale $S_t = \sum_{s \le t} x_s$ for

$$x_s = x_s(i) := \left(u_s^{\mathcal{R}}(i) - \ell_s^{\mathcal{R}}(i)\right).$$

We apply Corollary 31 to this supermartingale, taking

$$(a, b, M) = \left(\frac{4\log T}{\gamma_2} + 4\sqrt{\frac{\underline{L}^* \log T}{\gamma_2}}, \frac{\underline{L}^*}{\gamma_2}, \frac{1}{\gamma_2}\right).$$

For the filtration, we take the loss sequence as known from the start so that the only randomness is from the player's choices. Equivalently, we act as the observing adversary - note that S_t is still a supermartingale with respect to this filtration. Crucially, this means the conditional variance is bounded by $W_t \leq \frac{\ell_t^{\mathcal{R}}(i)}{\gamma_2}$. Therefore we have $V_t \leq \frac{L_t^{\mathcal{R}}(i)}{\gamma_2}$. We also note that with these parameters we have

$$e^{-\frac{a^2}{2b+Ma}} \leq e^{-\frac{a^2}{4b}} + e^{-\frac{a}{2M}} \leq \frac{1}{T^2} + \frac{1}{T^2} = \frac{2}{T^2}.$$

Therefore, Freedman's inequality tells us that with probability $1 - \frac{2}{T^2}$, for all t with $L_t^{\mathcal{R}}(i) \leq \underline{L}^*$ we have

$$S_t \leq a = \frac{4\log T}{\gamma_2} + 4\sqrt{\frac{\underline{L}^* \log T}{\gamma_2}}$$

and hence

$$U_t^{\mathcal{R}}(i) \leq L_t^{\mathcal{R}}(i) + \frac{4\log T}{\gamma_2} + 4\sqrt{\frac{\underline{L}^* \log T}{\gamma_2}} \leq \underline{L}^* + \frac{4\log T}{\gamma_2} + 4\sqrt{\frac{\underline{L}^* \log T}{\gamma_2}} \leq 2\underline{L}^* + \frac{8\log T}{\gamma_2}.$$

■

Proof of Lemma 33:

Similarly to the rare loss upper bound, we define an estimator for $L_t^{\mathcal{C}}(i)$:

$$U_t^{\mathcal{C}}(i) := \sum_{s\leq t} \frac{\ell_s^{\mathcal{C}}(i) \cdot 1_{i_s=i}}{\hat{p}_s(i)}.$$

We will again apply Freedman's inequality from the point of view of the adversary, this time to the martingale sequence $S_t = \sum_{s\leq t} x_s$ for

$$x_s = x_s(i) := \left(\frac{u_s^{\mathcal{C}}(i)}{\hat{p}_s(i)} - \ell_s^{\mathcal{C}}(i)\right).$$

We have $x_s \leq \frac{1}{\gamma_1} = M$ and $V_t \leq \frac{L_t^{\mathcal{C}}(i)}{\gamma_1}$. We use the parameters $b = \frac{\widetilde{L}}{\gamma_1}$ and $a = \lambda\sqrt{\frac{\widetilde{L}}{\gamma_1}}$. Using $\gamma \geq \frac{1}{\widetilde{L}}$ in the penultimate inequality and then $\lambda \geq 2$, we obtain the estimate:

$$e^{-\frac{a^2}{2b+Ma}} \leq e^{-\frac{a^2}{4b}} + e^{-\frac{a}{2M}} \leq e^{-\frac{\lambda^2}{4}} + e^{-\frac{\lambda^2\sqrt{\widetilde{L}\gamma_1}}{2}} \leq e^{-\frac{\lambda^2}{4}} + e^{-\frac{\lambda}{2}} \leq 2e^{-\frac{\lambda}{2}}.$$

Plugging into Freedman, we see that with probability at least $1 - 2e^{-\lambda/2}$, for all t with $L_t^{\mathcal{C}}(i) \leq \widetilde{L}$ we have

$$U_t^{\mathcal{C}}(i) \leq L_t^{\mathcal{C}}(i) + \lambda\sqrt{\frac{\widetilde{L}}{\gamma_1}}.$$

■

Now we use these lemmas to prove Theorem 15. In both halves, the main idea is that if something holds with high probability for any loss sequence, then the player must assign it high probability on average.

Proof of Theorem 15A:

Let E be the event that for all t with $L_t^{\mathcal{R}}(i) \leq \underline{L}^*$ we have

$$U_t^{\mathcal{R}}(i) \leq 2\underline{L}^* + \frac{8\log T}{\gamma_2}.$$

By Lemma 32 we have $\mathbb{P}[E] \geq 1 - \frac{2}{T^2}$ for any fixed loss sequence. The player does not know what the true loss sequence is, but his prior is a mixture of possible loss sequences, and so the player also assigns E a probability at least $1 - \frac{2}{T^2}$ at the start of the game. Let F denote the event that the player assigns E probability $\mathbb{P}_t[E] \geq 1 - \frac{1}{T}$ at all times during the game. Since probabilities are martingales, F has probability at least $1 - \frac{2}{T}$ by Doob's inequality.

Assume that F holds, so that $\mathbb{P}_t[E] \geq 1 - \frac{1}{T}$ at all times. After the first time τ that $U_t^{\mathcal{R}}(i) > 2\underline{L}^* + \frac{8\log T}{\gamma_2}$, as long as E holds we must have $L_t^{\mathcal{R}}(i) > \underline{L}^*$ and so $a^* \neq i$. Therefore, if F holds then after time τ we always have $p_t(i) \leq \frac{1}{T}$. So the expected number of additional times that arm i is pulled after this point is less than 1.

On the complementary event where F does not hold we simply observe that this event has probability at most $\frac{2}{T}$ and contributes loss at most T, therefore the expected loss from this event is at most 2.

To finish, we note that $\gamma_2 U_t^{\mathcal{R}}(i)$ is exactly the total loss paid by the player from arm i when i is rare. Therefore τ is the first time t which satisfies

$$\gamma_2 U_t^{\mathcal{R}}(i) > \gamma_2\left(2\underline{L}^* + \frac{8\log T}{\gamma_2}\right) = 2\gamma_2\underline{L}^* + 8\log T.$$

Assuming that F holds, this means that at time τ we have

$$\sum_{s\leq\tau} \ell_s^{\mathcal{R}}(i)1_{i_s=i} = \gamma_2 U_t^{\mathcal{R}}(i) \leq 2\gamma_2\underline{L}^* + 8\log(T) + 1.$$

It was just argued above that in this case i is pulled at most 1 additional time on average, and that the case when F is false contributes at most 2 loss of i in expectation. Therefore the total expected loss from i on rare rounds can be upper bounded by

$$2\gamma_2\underline{L}^* + 8\log T + 4.$$

■

Proof of Theorem 15B:

Let E_λ be the event that for all t with $L_t^{\mathcal{C}}(i) \leq \underline{L}^*$ we have

$$U_t^{\mathcal{C}}(i) \leq L_t^{\mathcal{C}}(i) + \lambda\sqrt{\frac{\underline{L}^*}{\gamma_1}}.$$

We apply Lemma 33 with $\widetilde{L} = \underline{L}^*$ which says that for $\lambda > 2$ we have $\mathbb{P}[E_\lambda] \geq 1 - e^{-\lambda/2}$. Let τ_λ be the first time that

$$U_t^{\mathcal{C}}(i) > \underline{L}^* + \lambda\sqrt{\frac{\underline{L}^*}{\gamma_1}}.$$

(If no such time exists, take $\tau_\lambda = +\infty$.) As before, note that at the beginning, the player must assign probability at least $1 - 2e^{-\lambda/2}$ to E since his prior is some mixture of loss sequences. By definition, if E holds then $i \neq a^*$ if time τ_λ is reached. Hence we see that $\mathbb{E}[p_{\tau_\lambda}(i)] \leq 2e^{-\lambda/2}$ by optional stopping since $U_t^{\mathcal{C}}(i)$ is computable (measurable) by the player. By Doob's maximal inequality, the probability that there exists $t \geq \tau_\lambda$ with $p_t(i) > \gamma_1$ is at most

$$\frac{2e^{-\lambda/2}}{\gamma_1} = 2e^{-\frac{\lambda - 2\log(1/\gamma_1)}{2}}.$$

Now, let λ^* be such that $U_t^{\mathcal{C}}(i) = \underline{L}^* + \lambda^*\sqrt{\frac{\underline{L}^*}{\gamma_1}}$ at the last time t when $p_t(i) > \gamma_1$. We have just shown an upper bound on the probability that there exists t with both

$$U_t^{\mathcal{C}}(i) > \underline{L}^* + \lambda\sqrt{\frac{\underline{L}^*}{\gamma_1}} \quad \text{and} \quad p_t(i) > \gamma_1.$$

So turning it the other way around, we see that

$$\mathbb{P}[\lambda^* > \lambda] \leq 2e^{-\frac{\lambda - 2\log(1/\gamma_1)}{2}}.$$

In other words, λ^* has tail bounded above by an exponential random variable with half-life $2\log(2)$ starting at $2\log(1/\gamma_1) + 2\log(2)$, and therefore

$$\mathbb{E}[\lambda^*] \leq 2\log(1/\gamma_1) + 10.$$

However, we always have $U_T^{\mathcal{C}}(i) = \underline{L}^* + \lambda^*\sqrt{\frac{\underline{L}^*}{\gamma_1}}$ since after the last time t with $p_t(i) > \gamma_1$, the value of $U_t^{\mathcal{C}}(i)$ cannot change. Recall also that $U_T^{\mathcal{C}}(i)$ is an unbiased estimator for $L_T^{\mathcal{C}}(i)$. Combining, we obtain:

$$\mathbb{E}[L_T^{\mathcal{C}}(i)] = \mathbb{E}[U_T^{\mathcal{C}}(i)] = \underline{L}^* + \mathbb{E}[\lambda^*]\sqrt{\frac{\underline{L}^*}{\gamma_1}} \leq \underline{L}^* + 2\left(\log\left(\frac{1}{\gamma_1}\right) + 10\right)\sqrt{\frac{\underline{L}^*}{\gamma_1}}.$$

■

A.3. Bandits with Better Mirror Maps

We begin by giving a re-interpretation of the work of (ZL19) for our setting. Call a function $f : [0,1] \to \mathbb{R}^+$ *admissible* when it satisfies:

1. $f'(x) \leq 0$ for all x.

2. $f''(x) \geq 0$ for all x.

3. $f'''(x) \leq 0$ for all x.

We then consider the potential function $F(p_t) = \sum_i f(p_t(i))$ for a probability vector p_t. The admissible functions we will consider are $f(x) = x\log(x)$ (negentropy), $f(x) = -x^{1/2}$ (negative Tsallis entropy), and $f(x) = -\log(Tx+1)$ (log barrier). We define $Max(F) = \max_{p\in\Delta_d} F(p)$ and $Min(F) = \min_{p\in\Delta_d} F(p)$ where Δ_d is the simplex of d-dimensional probability vectors. Note that convexity easily implies $Max(F) = F(1)+(d-1)F(0)$ and $Min(F) = dF(1/d)$. We set $diam(F) = Max(F) - Min(F)$.

The key point is that for any admissible f, we have:

Proposition 35

$$\mathbb{E}_t[F(p_{t+1}) - F(p_t)] \geq \sum_i \hat{p}_t(i)p_t(i)^2 f''(p_t(i)) \frac{\left(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right)_+^2}{2\bar{\ell}_t(i)}.$$

The proof is a one-sided quadratic estimate for f as in our reverse-chi-squared estimate.

Proof of Proposition 35:

We define (as in the proof of Lemma 13) $\widetilde{\ell}_t(i)$ to be a $\{0,1\}$ variable with mean $\bar{\ell}_t(i)$ and conditionally independent of everything else. We note that

$$\mathbb{P}[a^* = i|\widetilde{\ell}_t(i) = 1] = p_t(i|\widetilde{\ell}_t(i) = 1) = \frac{p_t(i)\bar{\ell}_t(i,i)}{\bar{\ell}_t(i)}$$

by Bayes rule. Therefore

$$\bar{\ell}_t(i,i) = \frac{p_t(i|\widetilde{\ell}_t(i) = 1)\bar{\ell}_t(i)}{p_t(i)}$$
$$\implies \bar{\ell}_t(i) - \bar{\ell}_t(i,i) = \bar{\ell}_t(i)\left(\frac{p_t(i) - p_t(i|\widetilde{\ell}_t(i) = 1)}{p_t(i)}\right).$$

As a result, we have

$$\sum_i p_t(i)^3 f''(p_t(i)) \frac{\left(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right)_+^2}{\bar{\ell}_t(i)} = \sum_i p_t(i)\bar{\ell}_t(i) f''(p_t(i)) \left(p_t(i) - p_t(i|\widetilde{\ell}_t(i) = 1)\right)_+^2.$$

Now, for any a, b we have by admissibility that

$$f(b) - f(a) \geq f'(a)(b-a) + \frac{f''(a)}{2}(a-b)_+^2.$$

Indeed, as in the proof of Lemma 8, $f(b)$ is convex on $b \geq a$ and $\frac{f''(a)}{2}$-strongly convex on $b \leq a$. Therefore:

$$f(p_t(i|\ell_t(i)) - f(p_t(i)) \geq f'(p_t(i))(p_t(i|\ell_t(i)) - p_t(i)) + \frac{f''(p_t(i))}{2}(p_t(i) - p_t(i|\ell_t(i)))_+^2$$

.

Now we take the expectation over $\ell_t(i)$ and note that $\mathbb{E}[p_t(i|\ell_t(i))] = p_t(i)$ by the martingale property. We compute:

$$\begin{aligned}\mathbb{E}_t[f(p_t(i|\ell_t(i)) - f(p_t(i))] \geq & \mathbb{E}\left[\frac{f''(p_t(i))}{2}(p_t(i) - p_t(i|\ell_t(i)))_+^2\right] \\ \geq & \frac{f''(p_t(i))}{2}\bar{\ell}_t(i)(p_t(i) - p_t(i|\widetilde{\ell}_t(i) = 1))_+^2.\end{aligned}$$

Hence

$$\mathbb{E}\left[\sum_i \hat{p}_t(i)\bar{\ell}_t(i)f''(p_t(i))\left(p_t(i) - p_t(i|\widetilde{\ell}_t(i) = 1)\right)_+^2\right] \leq 2\mathbb{E}\left[\sum_i \hat{p}_t(i)\left(f(p_t(i|\ell_t(i))) - f(p_t(i))\right)\right].$$

Now note that $f'' > 0$ implies that $f(X)$ is a submartingale for any martingale X, therefore

$$\begin{aligned}\mathbb{E}[F(p_{t+1}(i)) - F(p_t(i))] =& \sum_{i,j}\hat{p}_t(j)\mathbb{E}[f(p_t(i|\ell_t(j))) - f(p_t(j))] \\ \geq & \sum_i \hat{p}_t(i)\mathbb{E}[f(p_t(i|\ell_t(i))) - f(p_t(i))] \\ \geq & \sum_i \hat{p}_t(i)\bar{\ell}_t(i)\frac{f''(p_t(i))}{2}(p_t(i) - p_t(i|\widetilde{\ell}_t(i) = 1))_+^2 \\ =& \sum_i \hat{p}_t(i)p_t(i)^2 f''(p_t(i))\left(\frac{\left(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right)_+^2}{2\bar{\ell}_t(i)}\right).\end{aligned}$$

■

Cauchy-Schwarz and the rare/common decomposition gives the corollary:

Corollary 36

For any admissible f, and $\mathcal{C}_t, \mathcal{R}_t$ generated by a parameter $\gamma > 0$, Thompson Sampling satisfies

$$\mathbb{E}[R_T] \leq \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\ell_t(i)\right] + \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)(\ell_t(i) - \ell_t(i,i))\right]$$

and the two terms are bounded by

$$\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\ell_t(i)\right] \leq \min\left(\gamma T, d\cdot(2\gamma\underline{L}^* + 8\log(T) + 4)\right)$$

$$\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)(\ell_t(i) - \ell_t(i,i))\right] \leq \sqrt{2(Max(F) - F(p_1))\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}.$$

Proof of Corollary 36:

As in the proof of Theorem 16 we have

$$\mathbb{E}[R_T] \leq \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] + \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)\cdot(\bar{\ell}_t(i)-\bar{\ell}_t(i,i))_+\right].$$

For the first term, the upper bound of γ is immediate while Theorem 15 says

$$\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] \leq d\cdot(2\gamma\underline{L}^* + 8\log(T)+4).$$

For the second term, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)\cdot(\bar{\ell}_t(i)-\bar{\ell}_t(i,i))_+\right]\\
\leq&\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t} p_t(i)^3 f''(p_t(i))\frac{\left(\bar{\ell}_t(i)-\bar{\ell}_t(i,i)\right)_+^2}{\bar{\ell}_t(i)}}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}\\
\leq&\sqrt{\mathbb{E}\sum_{i,t} p_t(i)^3 f''(p_t(i))\frac{\left(\bar{\ell}_t(i)-\bar{\ell}_t(i,i)\right)_+^2}{\bar{\ell}_t(i)}}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}\\
\leq&\sqrt{2\sum_t \mathbb{E}_t[F(p_{t+1})-F(p_t)]}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}\\
\leq&\sqrt{2\cdot\mathbb{E}[F(p_T)-F(p_1)]}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}\\
\leq&\sqrt{2\cdot(Max(F)-F(p_1))\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}.
\end{aligned}$$

■

Now we can prove the refined bandit estimates.

Proof of Theorem 18:

We apply Corollary 36 with $f(x) = -x^\alpha$. Then $f''(x) = \alpha(1-\alpha)x^{\alpha-2}$ and $Max(F) = -1$, $Min(F) = -d^{1-\alpha}$. So the resulting bound is

$$
\begin{aligned}
\mathbb{E}[R_T] \leq & d \cdot (2\gamma \underline{L}^* + 8\log(T) + 4) + \sqrt{2(-1 + \sum_i p_1(i)^\alpha)} \sqrt{\mathbb{E} \sum_{(i,t): i \in \mathcal{C}_t} \frac{\bar{\ell}_t(i) p_t(i)^{1-\alpha}}{\alpha(1-\alpha)}} \\
\leq & d \cdot (2\gamma \underline{L}^* + 8\log(T) + 4) + O(\sqrt{H_\alpha(p_1)}) \left(\mathbb{E} \sum_{(i,t): i \in \mathcal{C}_t} \bar{\ell}_t(i) \right)^{\alpha/2} \left(\mathbb{E} \sum_{(i,t): i \in \mathcal{C}_t} \bar{\ell}_t(i) p_t(i) \right)^{(1-\alpha)/2} \\
\leq & d \cdot (2\gamma \underline{L}^* + 8\log(T) + 4) + O(\sqrt{H_\alpha(p_1)}) \left(\underline{L}^* + 2 \left(\log\left(\frac{1}{\gamma}\right) + 10 \right) \sqrt{\frac{\underline{L}^*}{\gamma}} \right)^{\alpha/2} (\mathbb{E}[L_T])^{(1-\alpha)/2}
\end{aligned}
$$

Taking $\gamma = \frac{\log^2(\underline{L}^*)}{\underline{L}^*}$ and assuming $\mathbb{E}[R_T] \geq 0$ (else any regret statement is vacuous) we get

$$
\begin{aligned}
\mathbb{E}[R_T] \leq & d \cdot (2\log^2(\underline{L}^*) + 8\log(T) + 4) + O(\sqrt{H_\alpha(p_1)}) \, (d\underline{L}^*)^{\alpha/2} \, (\mathbb{E}[L_T])^{(1-\alpha)/2} \\
\leq & d \cdot (2\log^2(\underline{L}^*) + 8\log(T) + 4) + O(\sqrt{H_\alpha(p_1)}) \cdot d^{\alpha/2} \, (\underline{L}^* + \mathbb{E}[R_T])^{1/2}.
\end{aligned}
$$

We apply the Lemma 37 below with:

- $R = \mathbb{E}[R_T]$
- $X = d \cdot (2\gamma \underline{L}^* + 8\log(T) + 4)$
- $Y = O(\sqrt{H_\alpha(p_1) d^\alpha})$
- $Z = \underline{L}^*$.

This gives the final regret bound

$$
\mathbb{E}[R_T] = O\left(\sqrt{H_\alpha(p_1) d^\alpha \underline{L}^*} + H_\alpha(p_1) d^\alpha + d\log^2(\underline{L}^*) + d\log(T) \right).
$$

■

Lemma 37

We have the general implication

$$
R \leq X + Y\sqrt{Z + R} \implies R \leq X + Y^2 + Y\sqrt{Z}.
$$

Proof

$$
\begin{aligned}
& R \leq X + Y\sqrt{Z + R} \\
\implies & R^2 - 2RX + X^2 \leq Y^2 Z + Y^2 R \\
\implies & R^2 - (2X + Y^2) R \leq Y^2 Z - X^2 \\
\implies & \left(R - \left(X + \frac{Y^2}{2} \right) \right)^2 \leq \frac{Y^4}{4} + Y^2 Z - X^2 \\
\implies & R \leq X + \frac{Y^2}{2} + \sqrt{\frac{Y^4}{4} + Y^2 Z - X^2} \leq X + Y^2 + Y\sqrt{Z}.
\end{aligned}
$$

■

Proof of Theorem 19:

We apply Corollary 36 with $f(x) = -\log(Tx+1)$. We have $diam(F) \approx d\log(T)$ and $f''(x) = \frac{1}{(x+T^{-1})^2}$. We also define $\mathcal{C}_t, \mathcal{R}_t$ using $\gamma = T^{-1}$. The $\mathcal{R}_t$ part of the bound in the Corollary is at most $\gamma T = 1$ so it remains to estimate the $\mathcal{C}_t$ sum.

To do this we observe that for $p_t(i) \geq \gamma = \frac{1}{T}$ we have

$$f''(p_t(i))^{-1} = (p_t(i) + T^{-1})^2 \leq p_t(i)^2 + 3p_t(i)T^{-1}.$$

Plugging in this estimate gives

$$\sum_{t,i:i\in\mathcal{C}_t} \frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))} \leq \sum_{t,i} \left(p_t(i)\bar{\ell}_t(i) + \frac{3\bar{\ell}_t(i)}{T}\right) \leq \mathbb{E}[L_T] + 6d.$$

Going back to the beginning and combining, we have shown

$$\mathbb{E}[R_T] = O\left(\sqrt{d\log(T)(\mathbb{E}[L_T] + d)} + 1\right) = O\left(\sqrt{d\mathbb{E}[L_T]\log(T)} + d\sqrt{\log(T)}\right).$$

Again, $a - b \leq \sqrt{ac}$ implies $a - b \leq \sqrt{bc} + c$ for non-negative a, b, c. We take

- $a = \mathbb{E}[L_T]$
- $b = \mathbb{E}[L^*] + O(d\sqrt{\log(T)})$
- $c = O(d\log(T))$

Therefore we obtain

$$\mathbb{E}[R_T] = O\left(\sqrt{d\mathbb{E}[L^*]\log(T)} + d\log(T)\right)$$

■

A.4. Semi-bandit Proof

We first illustrate why the naive extension of the $m = 1$ case fails to be tight in the semi-bandit case. Then we introduce the rank ordering of arms and explain how to define rare arms in this context. Finally, we prove Theorem 20.

A.4.1. Naive Analysis and Intuition

We let $a^* \in \mathcal{A}$ be the optimal set of m arms, and assume that it has loss at most $\underline{L}^*$. We let $\ell_t(i,i) = \mathbb{E}[\ell_t(i)|i \in a^*]$. Ignoring the issue of exactly how to assign arms as rare/common, we have

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t,i} p_t(i)(\ell_t(i) - \ell_t(i,i))\right] \leq \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\ell_t(i)\right] + \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))_+\right]$$

The first term is again small due to Theorem 15 and the second term can be estimated as

$$\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)\cdot(\bar{\ell}_t(i)-\bar{\ell}_t(i,i))_+\right] \le \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} \ell_t(i)\right]\cdot H^c(a^*).$$

The main difference is that now the coordinate entropy $H^c(a^*)$ can be as large as $\widetilde{O}(m)$. So the result is

$$\mathbb{E}[R_T] \le \widetilde{O}(\sqrt{H^c(a^*)d\underline{L}^*}) = \widetilde{O}(\sqrt{md\underline{L}^*}).$$

This argument is inefficient because it allows every arm to have loss $\underline{L}^*$ before becoming rare. However only j optimal arms can have loss more than $\frac{\underline{L}^*}{j}$. So although the coordinate entropy of a^* can be as large as $\widetilde{O}(m)$, the coordinate entropy on the arms with large loss so far is much smaller. This motivatives the rank ordering we introduce below.

A.4.2. Rare Arms and Rank Order

In this subsection we give two notions needed for the semi-bandit proof. First, analogously to our definition of rare and common arms in the bandit $m=1$ case, we split the action set $\mathcal{A}$ into rare and common arms. We start with an empty subset $\mathcal{R}_t=\emptyset\subseteq[d]$ of rare arms and grow it as follows. As long as there exists i so that $\mathbb{P}_t[(i\in a^*)$ and $(\mathcal{R}_t\cap a^*=\emptyset)]\le\gamma$, we add arm i to $\mathcal{R}_t$. Note that we must perform this procedure recursively because the actions have overlapping arms. The result is a time-varying partition of the arm set $[d]=\mathcal{R}_t\cup\mathcal{C}_t$ which at any time satisfies:

1. For any $i\in\mathcal{C}_t$ we have

$$p_t(i)\ge\mathbb{P}_t\left[(i\in a^*)\text{ and }(\mathcal{R}_t\cap a^*=\emptyset)\right]>\gamma.$$

2. For any $i\in\mathcal{R}_t$ we have

$$p_t(i)\le\mathbb{P}[a^*\not\subseteq\mathcal{C}_t]\le d\gamma.$$

As a result of this construction, for semi-bandit situations we will take $(\gamma_1,\gamma_2)=(\gamma,d\gamma)$ in applying Theorem 15.

The next step is to implement a rank ordering of the m coordinates. We let $a^*=\{a_1^*,a_2^*,\dots,a_m^*\}$ where $L_T(a_1^*)\ge L_T(a_2^*)\ge\cdots\ge L_T(a_m^*)$ and ties are broken arbitrarily. Crucially, we observe that

$$L_T(a_j^*)\le\frac{\underline{L}^*}{j}.$$

We further consider a general partition of $[m]$ into disjoint subsets $S_1,S_2,\dots,S_r$. Define $a^*_{S_k}=\{a^*_s : s\in S_k\}$. We will carry out an information theoretic argument on the events $\{i\in a^*_{S_k}\}$ and see that taking a dyadic partition allows a very efficient analysis. Our naive analysis corresponds to the trivial partition $S_1=[m]$.

A.4.3. Proof of Theorem 20

We first prove a suboptimal Shannon entropy version of the Theorem to guide the intuition for parts of the construction. The only difference is a multiplicative $\sqrt{\log(d)}$ in the main term.

Theorem 38 *The expected regret of Thompson Sampling in the semi-bandit case is*

$$O\left(\log(m)\sqrt{d\underline{L}^* \log(d)} + md^2 \log^2(\underline{L}^*) + d\log(T)\right).$$

Proof
First, we define:

$$p_t(i, S_k) := \mathbb{P}[i \in a^*_{S_k}]$$
$$\bar{\ell}(i, S_k) = \mathbb{E}[\ell_t(i) : i \in a_{S_k}]$$

We explain the changes from the $m = 1$ case, and then give the precise results. We again begin by bounding the regret by the total loss from rare arms plus the regret from common arms. We pick a threshold γ and apply the recursive procedure from the previous section. This means that Theorem 15 will apply with $(\gamma_1, \gamma_2) = (\gamma, d\gamma)$ for any γ. We set

$$(\gamma_1, \gamma_2) = \left(\frac{m\log^2(\underline{L}^*)}{\underline{L}^*}, \frac{md\log^2(\underline{L}^*)}{\underline{L}^*}\right).$$

Now for the analysis:

$$\mathbb{E}[R_T] \le \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] + \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right]$$
$$= \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] + \mathbb{E}\left[\sum_{(t,i,k):i\in\mathcal{C}_t} p_t(i,S_k)(\bar{\ell}_t(i) - \bar{\ell}_t(i,S_k))\right].$$

The first term is bounded by $O\left(md^2\log^2(\underline{L}^*) + d\log(T)\right)$ as a direct application of Theorem 15. For the second term we have by Cauchy-Schwarz that

$$\mathbb{E}\left[\sum_{(t,i,k):i\in\mathcal{C}_t} p_t(i,S_k)(\ell_t(i) - \ell_t(i,S_k))\right]$$

$$\le \sum_k \left(\mathbb{E}\left[\sum_{(t,i)} p_t(i)p_t(i,S_k)\left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i,S_k))_+^2}{\bar{\ell}_t(i)}\right)\right]^{1/2} \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} \frac{p_t(i,S_k)\ell_t(i)}{p_t(i)}\right]^{1/2}\right).$$

By Lemma 14 the first expectation can be estimated information theoretically by $H^c(a^*_{S_k})$:

$$\mathbb{E}\left[\sum_{(t,i)} p_t(i)p_t(i,S_k)\left(\frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i,S_k))_+^2}{\bar{\ell}_t(i)}\right)\right] \le 2\sum_t I_t^c[S_k] \le 2\cdot H^c(a^*_{S_k}).$$

Substituting this estimate, we have upper-bounded the common-arm regret term by

$$\sum_k \sqrt{2 \cdot H^c(a^*_{S_k})\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} \frac{p_t(i,S_k)\ell_t(i)}{p_t(i)}\right]}.$$

The key reason for introducing the sets S_k now appears, which is to give a separate estimate for the inner expectation. Let $s_k = \min(S_k)$. Observe that if $L_t(i) > \frac{\underline{L}^*}{s_k}$, then we cannot have $i \in a^*_{S_k}$ because $L_t(a^*_j) \le L_T(a^*_j) \le \frac{\underline{L}^*}{j} < L_t(i)$ for any $j \in S_k$. So roughly, for each fixed i the sum

$$\sum_{t:\, i\in\mathcal{C}_t} \frac{p_t(i,S_k)\bar{\ell}_t(i)}{p_t(i)}$$

will usually stop growing significantly once $L_t(i) > \frac{\underline{L}^*}{s_k}$ because $p_t(i,S_k)$ will be very small while $p_t(i) \ge \gamma$. Before this starts to happen, we have the simple estimate $\frac{p_t(i,S_k)}{p_t(i)} \le 1$. Therefore the sum should be bounded by approximately $\frac{\underline{L}^*}{s_k}$. In fact Lemma 39 below guarantees:

$$\mathbb{E}\left[\sum_{t:\, i\in\mathcal{C}_t} \frac{p_t(i,S_k)\bar{\ell}_t(i)}{p_t(i)}\right] \le \left(\frac{\underline{L}^*}{s_k} + O\left(\log\left(\frac{1}{\gamma_1}\right)\sqrt{\frac{\underline{L}^*}{s_k\gamma_1}}\right)\right).$$

Therefore using the fact that $H^c(a^*_{S_k}) = O(|S_k|\log(d))$ and multiplying by d to account for the d arms, we have an estimate of the common arm regret contribution of

$$O\left(\sum_k \sqrt{2d\log(d)|S_k|\underline{L}^*\left(\frac{1}{s_k} + O\left(\log\left(\frac{1}{\gamma_1}\right)\sqrt{\frac{1}{s_k\gamma_1\underline{L}^*}}\right)\right)}\right).$$

Since $\gamma_1 = \frac{m\log^2\underline{L}^*}{\underline{L}^*}$ we have

$$\log\left(\frac{1}{\gamma_1}\right)\sqrt{\frac{1}{s_k\gamma_1\underline{L}^*}} = O\left(\sqrt{\frac{1}{s_k m}}\right).$$

Substituting, and observing that

$$\sqrt{\frac{1}{s_k} + O\left(\sqrt{\frac{1}{s_k m}}\right)} = \sqrt{\frac{1}{s_k}} + O\left(\sqrt{\frac{1}{m}}\right)$$

we have that the common arm regret is at most

$$O\left(\sum_k \sqrt{2d\log(d)|S_k|\underline{L}^*\left(\frac{1}{s_k} + O\left(\sqrt{\frac{1}{s_k m}}\right)\right)}\right) = \sqrt{d\log(d)\underline{L}^*}\cdot O\left(\sum_k \sqrt{\frac{|S_k|}{s_k}}\right).$$

We are left with finding a partition $(S_1,\dots,S_r)$ that makes this last sum small. Taking the whole set $S_1 = [m]$ as in the naive analysis gives $\sqrt{m}$, while taking d singleton subsets $S_k = \{k\}$ gives $\sum_{k=1}^m k^{-1/2} =$

$\Theta(\sqrt{m})$. But taking a dyadic decomposition does much better! Letting $S_k = \{2^{k-1}, \dots, 2^k - 1\}$ gives a sum of

$$\sum_{k \le \lceil \log_2(m) \rceil} \sqrt{2} = O(\log m).$$

This yields a common arm regret estimate of

$$O(\log(m)\sqrt{d \log(d) \underline{L}^*}.$$

Combining with the estimate for rare arms, we have the claimed result.

■

Lemma 39 *Fix a subset $S_k \subseteq [m]$, let $s_k = \min(S_k)$, and let γ_1 be a constant satisfying $\frac{m}{\underline{L}^*} \le \gamma_1 \le \frac{1}{2}$. Thompson Sampling satisfies*

$$\mathbb{E}\left[\sum_{t:\, i \in \mathcal{C}_t} \frac{p_t(i, S_k)\bar{\ell}_t(i)}{p_t(i)}\right] \le \frac{\underline{L}^*}{s_k} + O\left(\log\left(\frac{1}{\gamma_1}\right)\sqrt{\frac{\underline{L}^*}{s_k \gamma_1}}\right).$$

Proof of Lemma 39: We first apply Lemma 33 with $\gamma_1 = \gamma$ and

$$\widetilde{L} = \frac{\underline{L}^*}{s_k}.$$

The lemma says that for $\lambda \ge 2$ and $\gamma_1 \ge \frac{s_k}{\underline{L}^*}$, with probability at least $1 - 2e^{-\lambda/2}$, for all t with $L_t^{\mathcal{C}}(i) \le \frac{\underline{L}^*}{s_k}$ we have

$$U_t^{\mathcal{C}}(i) \le L_t^{\mathcal{C}}(i) + \lambda\sqrt{\frac{\underline{L}^*}{s_k\gamma_1}} \le \frac{\underline{L}^*}{s_k} + \lambda\sqrt{\frac{\underline{L}^*}{s_k\gamma_1}}.$$

We note that since $p_t(i, S_k) \le p_t(i)$ and $\gamma \le p_t(i)$ we have

$$\mathbb{E}\left[\sum_{t:\, i \in \mathcal{C}_t} \frac{p_t(i, S_k)\bar{\ell}_t(i)}{p_t(i)}\right] \le A + 1 + \left(\frac{1}{\gamma}\right)\mathbb{E}\left[\sum_{t:\, i \in \mathcal{C}_t, L_t^{\mathcal{C}}(i) \ge A} p_t(i, S_k)\bar{\ell}_t(i)\right].$$

for any A. We rewrite the latter expectation, then essentially rewrite it again as a Riemann-Stieltjes integral (where $p_t(i, S_k) = p_{\lfloor t \rfloor}(i, S_k)$ for any positive real t):

$$\mathbb{E}\left[\sum_{t:\, i \in \mathcal{C}_t, L_t^{\mathcal{C}}(i) \ge A} p_t(i, S_k)\bar{\ell}_t(i)\right] = \mathbb{E}\left[\sum_{L_t^{\mathcal{C}}(i) \ge A} p_t(i, S_k)\ell_t^{\mathcal{C}}(i)\right]$$
$$\le \mathbb{E}\left[\int_A^\infty p_t(i, S_k)\, dL_t^{\mathcal{C}}(i)\right].$$

Define τ_x to be the first value of t satisfying

$$L_t^{\mathcal{C}}(i) \ge x$$

and set $\tau_x = \infty$ if $L_T^{\mathcal{C}}(i) < x$. Since losses are bounded by 1 we always have $t \geq \tau_{L_t^{\mathcal{C}}(i)-1}$. Therefore, changing variables from t to $L_t^{\mathcal{C}}(i)$ gives

$$\mathbb{E}\left[\int_A^\infty p_t(i,S_k)dL_t^{\mathcal{C}}(i)\right] \leq \mathbb{E}\left[\int_A^\infty \max_{t\geq\tau_{x-1}}\left(p_t(i,S_k)\right)\cdot 1_{\tau_x<\infty}dx\right]$$

$$\leq \mathbb{E}\left[\int_A^\infty \max_{t\geq\tau_{x-1}}\left(p_t(i,S_k)\right)\cdot 1_{\tau_{x-1}<\infty}dx\right] \leq 1+\mathbb{E}\left[\int_A^\infty \max_{t\geq\tau_x}(p_t(i,S_k))\cdot 1_{\tau_x<\infty}dx\right]$$

Translating the result of Lemma 33 shows that when $x = \frac{L^*}{s_k} + \lambda\sqrt{\frac{L^*}{s_k\gamma_1}}$ for $\lambda > 2$ we have

$$\mathbb{E}[p_{\tau_x}(i,S_k)1_{\tau_x<\infty}] \leq 2e^{-\lambda/2}.$$

Now, the average maximum of a martingale started at p and bounded in $[0,1]$ is seen by Doob's inequality to be at most $p(1-\log p)$. Therefore

$$\mathbb{E}[\max_{t\geq\tau_x} p_t(i,S_k)1_{\tau_x<\infty}|\mathcal{F}_{\tau_x}] \leq p_{\tau_x}(i,S_k)\cdot(1-\log\left(p_t(i,S_k)\right))\cdot 1_{\tau_x<\infty}.$$

The function $f(x) = x(1-\log x)$ is concave, so by Jensen's we have

$$\mathbb{E}[\max_{t\geq\tau_x} p_t(i,S_k)1_{\tau_x<\infty}] \leq 2(1+\lambda/2)e^{-\lambda/2} = (\lambda+2)e^{-\lambda/2}$$

where λ is such that $x = \frac{L^*}{s_k} + \lambda\sqrt{\frac{L^*}{s_k\gamma_1}}$.

So taking $A = \frac{L^*}{s_k} + 10\log\left(\frac{1}{\gamma}\right)\sqrt{\frac{L^*}{s_k\gamma_1}}$ and changing variables to integrate over λ gives an estimate of

$$A + \sqrt{\frac{L^*}{s_k\gamma_1}}\int_{10\log(1/\gamma)}^\infty (\lambda+2)e^{-\lambda/2}d\lambda.$$

This integral is bounded since $\gamma_1 \leq \frac{1}{2}$ (and also $10\log(1/\gamma) \geq 2$, which is needed since only $\lambda > 2$ is allowed in Lemma 33) so we get a bound of

$$\frac{L^*}{s_k} + O\left(\log\left(\frac{1}{\gamma_1}\right)\sqrt{\frac{L^*}{s_k\gamma_1}}\right).$$

■

We now explain how to modify the above analysis to use the Tsallis entropy and the log barrier instead of the Shannon entropy. Because the prior-dependence now requires the rank ordering to state, we suppress it for simplicity and obtain bounds depending on $diam(F)$. Similar to Corollary 36 we have the following lemma.

Lemma 40

For any admissible f, and $\mathcal{C}_t, \mathcal{R}_t$ generated by (γ_1, γ_2), Thompson Sampling for a semibandit problem satisfies

$$\mathbb{E}[R_T] \leq \mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] + \mathbb{E}\left[\sum_{(t,i,k):i\in\mathcal{C}_t} p_t(i,S_k)(\bar{\ell}_t(i)-\bar{\ell}_t(i,S_k))\right]$$

where

$$\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{R}_t} p_t(i)\bar{\ell}_t(i)\right] \le \min\left(\gamma_2 T, md^2\log^2(\underline{L}^*) + d\log(T)\right)$$

$$\mathbb{E}\left[\sum_{(t,i,k):i\in\mathcal{C}_t} p_t(i,S_k)(\bar{\ell}_t(i)-\bar{\ell}_t(i,S_k))\right] \le \sum_k \sqrt{2\cdot diam_{|S_k|}(F)\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}.$$

Here $diam_j(F)$ *is the diameter of* $F=\sum_{i=1}^d f(x_i)$ *restricted to* $\{x\in[0,1]^d, \sum_{i=1}^d x_i = j\}$.

Proof

Everything is clear except the final statement. Denote $p_t(S_k)$ the vector with $(p_t(S_k))_i = \mathbb{P}^t[i\in S_k]$. Then the calculation (whose justification is completely analogous to the $m=1$ setting) goes:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i,S_k)\cdot(\bar{\ell}_t(i)-\bar{\ell}_t(i,S_k))_+\right]\\
\le&\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t} p_t(i)p_t(i,S_k)^2 f''(p_t(i,S_k))\frac{\left(\bar{\ell}_t(i)-\bar{\ell}_t(i,S_k)\right)_+^2}{\bar{\ell}_t(i)}}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}\\
\le&\sqrt{\mathbb{E}\sum_{i,t} p_t(i)p_t(i,S_k)^2 f''(p_t(i,S_k))\frac{\left(\bar{\ell}_t(i)-\bar{\ell}_t(i,S_k)\right)_+^2}{\bar{\ell}_t(i)}}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}\\
\le&\sqrt{2\sum_t \mathbb{E}_t[F(p_{t+1}(S_k))-F(p_t(S_k))]}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}\\
\le&\sqrt{2(F(p_T(S_k))-F(p_1(S_k)))}\cdot\sqrt{\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}\\
\le&\sqrt{2\cdot diam_{|S_k|}(F)\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}.
\end{aligned}$$

■

Proof of Theorem 20:

We apply the preceding lemma with $f(x)=-x^{1/2}$ and S_k the dyadic partition of $[m]$ and $(\gamma_1,\gamma_2) = (\frac{m\log^2(\underline{L}^*)}{\underline{L}^*}, \frac{md\log^2(\underline{L}^*)}{\underline{L}^*})$. Then $f''(x)=\frac{1}{4x^{3/2}}$. We have $diam_j(F)\le\sqrt{jd}$. Therefore the common arm regret is at most

$$\mathbb{E}\left[\sum_{(t,i,k):i\in\mathcal{C}_t} p_t(i,S_k)(\bar{\ell}_t(i)-\bar{\ell}_t(i,S_k))\right] \leq O\left(\sum_k \sqrt{2^{k/2}d^{1/2}\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i,S_k))}}\right)$$
$$\leq O\left(\sum_k \sqrt{2^{k/2}d^{1/2}\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{p_t(i,S_k)^{3/2}\bar{\ell}_t(i)}{p_t(i)}}\right).$$

Now from Cauchy-Schwarz and $p_t(i,S_k)\leq p_t(i)$ we have

$$\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{p_t(i,S_k)^{3/2}\ell_t(i)}{p_t(i)} \leq \left(\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{p_t(i,S_k)\ell_t(i)}{p_t(i)}\right)^{1/2}\cdot\left(\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}p_t(i,S_k)\ell_t(i)\right)^{1/2}.$$

By Lemma 39 (with an extra factor of d for summing over all arms) and using $\gamma_1=\frac{m\log^2(\underline{L}^*)}{\underline{L}^*}$ we obtain

$$\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{p_t(i,S_k)\ell_t(i)}{p_t(i)} = d\cdot O\left(\frac{\underline{L}^*}{2^k}+\log\left(\frac{1}{\gamma_1}\right)\sqrt{\frac{\underline{L}^*}{2^k\gamma_1}}\right) = O\left(\frac{d\underline{L}^*}{2^{k/2}}\right).$$

Also it is clear by definition that

$$\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}p_t(i,S_k)\ell_t(i)\leq\mathbb{E}[L_T].$$

Combining we have a common arm regret of at most

$$O\left(\sum_k\sqrt{2^{k/2}d^{1/2}\cdot\sqrt{\frac{dL^*}{2^k}}\cdot\mathbb{E}[L^T]}\right) = O\left(\sum_k\sqrt{dL_T}\right) = O\left(\log(m)\sqrt{dL_T}\right).$$

Adding in the rare arm regret gives:

$$\mathbb{E}[R_T]\leq O\left(md^2\log^2(\underline{L}^*)+d\log(T)+\log(m)\sqrt{d\cdot(\underline{L}^*+\mathbb{E}[R_T])}\right)$$

Now we apply Lemma 37 with:

- $R=\mathbb{E}[R_T]$
- $X=O(md^2\log^2(\underline{L}^*)+d\log(T))$
- $Y=O(\log(m)\sqrt{d})$
- $Z=\underline{L}^*$

The result is as claimed:

$$\mathbb{E}[R_T]\leq O\left(\log(m)\sqrt{d\underline{L}^*}+md^2\log^2(\underline{L}^*)+d\log(T)\right)$$

■

We now prove the analogous result with the log barrier.

Proof of Theorem 21:

We apply the lemma with $f(x) = -\log(Tx+1)$ and $(\gamma_1, \gamma_2) = (\frac{1}{T}, \frac{d}{T})$ with no partitioning scheme, i.e. $S_1 = [m]$. Then $f''(x)^{-1} = (x + T^{-1})^2 \leq x^2 + \frac{3x}{T}$ when $x \geq T^{-1}$. We have $diam_j(F) \leq d\log(T)$. Therefore the common arm regret is at most

$$\begin{aligned}
\mathbb{E}\left[\sum_{(t,i):i\in\mathcal{C}_t} p_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right] \leq & O\left(\sqrt{d\log(T)\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}\frac{\bar{\ell}_t(i)}{p_t(i)f''(p_t(i))}}\right) \\
\leq & O\left(\sqrt{d\log(T)\cdot\mathbb{E}\sum_{(i,t):i\in\mathcal{C}_t}(p_t(i)+3T^{-1})\bar{\ell}_t(i)}\right) \\
\leq & O\left(\sqrt{d\log(T)\cdot(\mathbb{E}[L_T]+3d)}\right) \\
\leq & O\left(\sqrt{d\log(T)\cdot\mathbb{E}[L_T]} + d\sqrt{\log(T)}\right)
\end{aligned}$$

The rare arm regret is at most d which is absorbed in the $O(d\sqrt{\log(T)})$. The result now follows from the same computation as the end of the proof of Theorem 19.

■

A.5. Thresholded Thompson Sampling

Here we prove Theorems 22 and 23. First we precisely define thresholded Thompson Sampling. For some parameter $\gamma > 0$, we have already described how to generate a partition $[d] = \mathcal{R}_t \cup \mathcal{C}_t$ into common and rare actions. In the bandit case, we define Thompson Sampling thresholded at γ as a Bayesian algorithm playing from the following distribution:

$$\hat{p}_t(i) = \begin{cases} 0 & \text{if } i \in \mathcal{R}_t \\ \frac{p_t(i)}{1-\sum_{j\in\mathcal{R}_t} p_t(j)} & \text{if } i \in \mathcal{C}_t. \end{cases}.$$

In the semi-bandit case, we use the following definition which generalizes the above:

$$\hat{p}_t(a_t = a) = \begin{cases} 0 & \text{if } a \cap \mathcal{R}_t \neq \emptyset \\ \frac{p_t(a)}{1-\sum_{a'\cap\mathcal{R}_t\neq\emptyset} p_t(a')} & \text{if } a \cap \mathcal{R}_t = \emptyset. \end{cases}$$

Note that in the semi-bandit case, for $i \in \mathcal{C}_t$ we may have $\hat{p}_t(i) < p_t(i)$. However we always have $\hat{p}_t(i) \geq \gamma$, so Theorem 15 still applies.

Theorem 22

Suppose that the best action almost surely has total loss at most $\underline{L}^*$. Thompson Sampling for bandit feedback, thresholded with $\gamma = \frac{\log^2(\underline{L}^*)}{\underline{L}^*} \leq \frac{1}{2d}$, has expected regret

$$\mathbb{E}[R_T] = O\left(\sqrt{d\underline{L}^*} + d\log^2(\underline{L}^*)\right)$$

Theorem 23

Suppose that the best combinatorial action almost surely has total loss at most $\underline{L}^*$. Thompson Sampling for semi-bandit feedback, thresholded with $\gamma = \frac{m\log^2(\underline{L}^*)}{\underline{L}^*} \leq \frac{1}{2d}$, has expected regret

$$\mathbb{E}[R_T] = O\left(\log(m)\sqrt{d\underline{L}^*} + md\log^2(\underline{L}^*)\right).$$

Proof of Theorem 22:

Note that we always have

$$\hat{p}_t(i) \leq \frac{p_t(i)}{1-\gamma d}.$$

We therefore have the calculation

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{i,t} \hat{p}_t(i)\bar{\ell}_t(i) - p_t(i)\ell_t(i,i)\right]$$

$$= \mathbb{E}\left[(\hat{p}_t(i) - p_t(i))\bar{\ell}_t(i,i)\right] + \mathbb{E}\left[\hat{p}_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right]$$

$$\leq \left(\frac{\gamma d}{1-\gamma d}\right) \cdot \mathbb{E}\left[\sum_{i,t} p_t(i)\bar{\ell}_t(i,i)\right] + \mathbb{E}\left[\sum_{(i,t):i\in\mathcal{C}_t} \hat{p}_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right].$$

The former expectation is at most $\underline{L}^*$ while the second can be bounded in the same way as the non-thresholded results. Indeed, since $\hat{p}_t(i)/p_t(i) \approx 1$ whenever $i \in \mathcal{C}_t$ there is no issue in adapting any of the proofs.

The result is that we get the same common arm regret as in the proof of Theorem 18 but with only an additive $d\log^2(\underline{L}^*)$ from the rare arms. From here it is easy to conclude as in that proof.

■

Proof of Theorem 23:

If we threshold at γ, we remove at most γd probability of actions, so we still have

$$\hat{p}_t(i) \leq \frac{p_t(i)}{1-\gamma d}.$$

The corresponding calculation is

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{i,t} \hat{p}_t(i)\bar{\ell}_t(i) - p_t(i)\ell_t(i,i)\right]$$

$$= \mathbb{E}\left[(\hat{p}_t(i) - p_t(i))\bar{\ell}_t(i,i)\right] + \mathbb{E}\left[\hat{p}_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right]$$

$$\leq \left(\frac{\gamma d}{1-\gamma d}\right) \cdot \mathbb{E}\left[\sum_{i,t} p_t(i)\bar{\ell}_t(i,i)\right] + \mathbb{E}\left[\sum_{(i,t):i\in\mathcal{C}_t} \hat{p}_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right].$$

$$\leq \left(\frac{\gamma d}{1-\gamma d}\right)\underline{L}^* + \sum_k \mathbb{E}\left[\sum_{(i,t):i\in\mathcal{C}_t} \frac{\hat{p}_t(i)p_t(i,S_k)}{p_t(i)}(\bar{\ell}_t(i) - \bar{\ell}_t(i,S_k))\right].$$

We take the sets S_k to be the same dyadic partition of $[m]$ as in the non-thresholded case. Thresholding at $\gamma = \frac{m\log^2(\underline{L}^*)}{\underline{L}^*}$ makes the first term $md\log^2(\underline{L}^*)$, and the second term again satisfies the same estimates as in the non-thresholded case up to a $1+\gamma d = O(1)$ multiplicative factor. So again we get the same estimates as in the proof of Theorem 20 but with no rare arm terms, which easily gives the result.

■

A.6. Proofs of Negative Results

A.6.1. Thompson Sampling Does Not Satisfy High-Probability Regret Bounds

It is natural to ask for high-probability regret bounds. Here we point out that against a worst-case prior, Thompson Sampling does not satisfy any high-probability regret bound even with full feedback.

Theorem 27

There exist prior distributions for which Thompson Sampling achieves $\Omega(T)$ regret a constant fraction of the time, with either full or bandit feedback.

Proof

We construct such a prior distribution with 2 arms. First, arm 1 sees loss 1 for each the first $T/3$ rounds while arm 2 sees none. Then with probability 50%, both arms see no more loss, while with probability 50%, arm 2 sees loss 1 for the last $2T/3$ rounds but arm 1 still sees no more loss.

In this example, Thompson Sampling will pick arm 1 half the time for the first $T/3$ rounds, hence has a 50% chance to have a linear regret from the case that there no further loss.

■

A.6.2. Thompson Sampling Does Not Achieve Full T-independence

Our first-order regret bounds for non-thresholded Thompson Sampling in the (semi)bandit cases had $d\log(T)$ terms. Here we show that mild T-dependence is inherent to the algorithm.

Theorem 28

There exist prior distributions against which Thompson Sampling achieves $\Omega(d\underline{L}^*)$ expected regret for very large T with bandit feedback, even given the value $\underline{L}^*$.

Proof

An example prior distribution for at least 3 arms is as follows. First pick a uniformly random "good" arm. For the others, flip a coin to decide if they are "bad" or "terrible".

We insist that the good arm have loss 0 on the first round and the other arms have loss 1. For the good arm, every subsequent loss is a fair coin flip until the total loss reaches $\underline{L}^*$. For each bad arm we do the same but stop when the loss reaches $\underline{L}^*+1$. The terrible arms receive fair coin flip losses forever.

Assume that the player does not pick the good arm on the first time-step. We claim that given infinite time, the player will pay loss $\underline{L}^*+1$ on each terrible arm, which implies the desired result.

Indeed, suppose the player played a terrible arm i most recently at time $\hat{t}$, did not play i on the first round, and has observed loss less than $\underline{L}^*$ on this arm. Then we claim that the player's probability for arm i to be good must be bounded away from 0 for any fixed $\hat{t}$. Indeed, the initialization with i as the good arm and the good arm to be bad must be uniformly positive. The only Bayesian evidence for the truth over this alternative hypothesis is the player's observed losses on arm i. But these observations are inconclusive and since t is fixed their Bayes factor stays bounded.

Additionally, with probability 1 the player's probability assigned to the true arm configuration is bounded away from 0 uniformly in time. Indeed, being a martingale, if this were false then the probability would have to converge to 0. But the player's subjective probability of a true statement cannot converge to 0, because the player after an infinite amount of time would assign the true statement probability 0.

Since for fixed $\hat{t}$ the Bayes factor between the truth and the alternative is bounded, we see that this alternative with arm i as the good arm has probability bounded away from 0 uniformly in time.

We have just argued that Thompson Sampling with this prior will have a uniformly positive probability to play such an arm i until it plays i again. Thus, with probability $\frac{d-1}{d}$ (for the first arm not to be good), Thompson Sampling accumulates loss $\underline{L}^* + 1$ on every terrible arm except the first arm it plays. This results in $\Omega(d\underline{L}^*)$ regret.

■

Theorem 29

It is possible that Thompson Sampling achieves, with high probability, loss $\Omega(\sqrt{T})$ for a contextual bandit problem with $L^* = 0$ optimal loss, $K = 2$ cliques, and $d = O(\sqrt{T})$ total arms.

Proof

Set $S = \sqrt{T}$ and fix $d \geq 2S$. Form S distinct *small cliques*, with random but disjoint sets of $\frac{d}{2S}$ arms each. Call these cliques $C_1, \dots, C_S$. Also generate independent random bits $b_1, \dots, b_S$. Then for $j \in \{0, 1, \dots, \sqrt{T} - 1\}$ make the loss b_i on the small clique C_j for rounds $\{jS + 1, \dots, (j+1)S\}$ and 0 on the complement $[d] \backslash C_j$. (So the feedback graphs are cliques on C_j and $[d] \backslash C_j$.) Finally, in the last round pick at random a single arm a_k with no loss so far and make the loss 0 for a_k and 1 for all other arms.

Then clearly $L^* = 0$ for this arm a_k. However Thompson Sampling will incur a constant expected loss for each clique C_j hence pays total expected cost $\Theta(S) = \Theta(\sqrt{T})$.

■