35 Thompson Sampling

"As all things come to an end, even this story, a day came at last when they were in sight of the country where Bilbo had been born and bred, where the shapes of the land and of the trees were as well known to him as his hands and toes." – Tolkien [1937].

Like Bilbo, as we near the end of the book we return to where it all began, to the first algorithm for bandits proposed by Thompson [1933]. The idea is a simple one. Before the game starts the learner chooses a prior over a set of possible bandit environments. In each round the learner samples an environment from the posterior and acts according to the optimal action in that environment. Thompson only gave empirical evidence (calculated by hand) and focussed on Bernoulli bandits with two arms. Nowadays these limitations have been eliminated and theoretical guarantees have been proven demonstrating the approach is often close to optimal in a wide range of settings. Perhaps more importantly, the resulting algorithms are often quite practical both in terms of computation and empirical performance. The idea of sampling from the posterior and playing the optimal action is called **Thompson sampling** or **posterior sampling**.

The exploration in Thompson sampling comes from the randomization. If the posterior is poorly concentrated, then the fluctations in the samples are expected to be large and the policy will likely explore. On the other hand, as more data is collected the posterior concentrates towards the true environment and the rate of exploration decreases. We focus our attention on finite-armed stochastic bandits and linear stochastic bandits, but Thompson sampling has been extended to all kinds of models as explained in the bibliographic remarks.

Randomization is crucial for adversarial bandit algorithms and can be useful in stochastic settings (see Chapters 23 and 32 for examples). We should be wary, however, that there might be a price to pay by injecting variance into our algorithms. What is gained or lost by the randomization in Thompson sampling is still not clear, but we leave this cautionary note as a suggestion to the reader to think about some of the costs and benefits.

35.1 Finite-armed bandits

Recalling very briefly the notation from Section 34.5, let $K > 1$ be the number of arms and $(\mathcal{E}, \mathcal{G}, \mathbb{Q})$ be a probability space where $\mathcal{E}$ is a set of K-armed stochastic bandits and $\mathbb{Q}$ is the prior. For $\nu \in \mathcal{E}$ the distribution on the reward vector in each round is $\mathbb{P}_\nu$ on $(\mathbb{R}^K, \mathfrak{B}(\mathbb{R}^K))$. As usual we assume that $\mathbb{P}_\nu$ is a probability kernel from $(\mathcal{E}, \mathcal{G})$ to $(\mathbb{R}^K, \mathfrak{B}(\mathbb{R}^K))$. The reward vector in round t is $X_t \in \mathbb{R}^K$ and the learner observes X_{tA_t}. The posterior after t observations is a random measure $\mathbb{Q}_t$ on $(\mathcal{E}, \mathcal{G})$. The mean of the ith arm in bandit $\nu \in \mathcal{E}$ is $\mu_i(\nu)$.

1: **Input** K, $\mathcal{E}$ and prior $\mathbb{Q}$
2: **for** $t = 1, 2, \dots n$ **do**
3: Compute posterior $\mathbb{Q}_{t-1}$ based on observed data
4: Sample $\nu_t \sim \mathbb{Q}_{t-1}$
5: Choose $A_t = \operatorname{argmax}_{i \in [K]} \mu_i(\nu_t)$
6: **end for**

Bayesian analysis

Thompson sampling has been analyzed in both the frequentist and the Bayesian settings. We start with the latter where the result requires almost no assumptions on the prior. In fact, after one small observation about Thompson sampling, the analysis is almost the same as that of UCB.

THEOREM 35.1 *Let $\mathcal{E}$ a set of 1-subgaussian bandits with K arms and mean rewards bounded in $[0, 1]$ and $\mathbb{Q}$ be a measure on $(\mathcal{E}, \mathcal{G})$ for some σ-algebra $\mathcal{G}$ and π be the policy of Thompson sampling with this prior. Then*

$$\operatorname{BR}_n(\pi, \mathbb{Q}) \leq C \sqrt{Kn \log(n)} \,,$$

where $C > 0$ is a universal constant.

Proof Let $\mathbb{P}$ be the joint measure defined after Eq. (34.5) and ν, A_t and X_t be the coordinate projections given in Eq. (34.6). Expectations are taken with respect to $\mathbb{P}$. Abbreviate $\mu_i = \mu_i(\nu)$ and let $A^* = \operatorname{argmax}_{i \in [K]} \mu_i$ be the optimal arm, which depends on ν and is a random variable. For each $t \in [n]$ and $i \in [K]$ let

$$U_t(i) = \operatorname{clip}_{[0,1]} \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log(1/\delta)}{1 \vee T_i(t-1)}} \right) ,$$

where $\hat{\mu}_i(t-1)$ is the empirical estimate of the reward of arm i after $t-1$ rounds and we assume $\hat{\mu}_i(t-1) = 0$ if $T_i(t-1) = 0$. Let E be the event that for all $t \in [n]$ and $i \in [K]$,

$$|\hat{\mu}_i(t-1) - \mu_i| < \sqrt{\frac{2 \log(1/\delta)}{1 \vee T_i(t-1)}} \,.$$

In Exercise 35.1 we ask you to prove that $\mathbb{P}(E^c) \leq nK\delta$. Note that $U_t(i)$ is $\mathcal{F}_{t-1}$-measurable. The Bayesian regret is

$$\mathrm{BR}_n = \mathbb{E}\left[\sum_{t=1}^{n}(\mu_{A^*} - \mu_{A_t})\right] = \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}\left[\mu_{A^*} - \mu_{A_t} \mid \mathcal{F}_{t-1}\right]\right].$$

The key insight is to notice that the definition of Thompson sampling implies the conditional distributions of A^* and A_t given $\mathcal{F}_{t-1}$ are the same:

$$\mathbb{P}(A^* = \cdot \mid \mathcal{F}_{t-1}) = \mathbb{P}(A_t = \cdot \mid \mathcal{F}_{t-1}) \quad \text{a.s.}. \tag{35.1}$$

Using the previous display,

$$\begin{aligned}
\mathbb{E}\left[\mu_{A^*} - \mu_{A_t} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mu_{A^*} - U_t(A_t) + U_t(A_t) - \mu_{A_t} \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mu_{A^*} - U_t(A^*) + U_t(A_t) - \mu_{A_t} \mid \mathcal{F}_{t-1}\right] \quad \text{(Eq. (35.1))} \\
&= \mathbb{E}\left[\mu_{A^*} - U_t(A^*) \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[U_t(A_t) - \mu_{A_t} \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

Using the tower rule for expectation shows that

$$\mathrm{BR}_n = \mathbb{E}\left[\sum_{t=1}^{n}(\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^{n}(U_t(A_t) - \mu_{A_t})\right]. \tag{35.2}$$

On the event E^c the terms inside the expectation are bounded by $2n$ while on the event E the first sum is negative and the second is bounded by

$$\begin{aligned}
\mathbb{I}\{E\}\sum_{t=1}^{n}(U_t(A_t) - \mu_{A_t}) &= \mathbb{I}\{E\}\sum_{t=1}^{n}\sum_{i=1}^{K}\mathbb{I}\{A_t = i\}(U_t(i) - \mu_i) \\
&\leq \sum_{i=1}^{K}\sum_{t=1}^{n}\mathbb{I}\{A_t = i\}\sqrt{\frac{8\log(1/\delta)}{1 \vee T_i(t-1)}} \leq \sum_{i=1}^{K}\int_0^{T_i(n)}\sqrt{\frac{8\log(1/\delta)}{s}}ds \\
&= \sum_{i=1}^{K}\sqrt{32T_i(n)\log(1/\delta)} \leq \sqrt{32nK\log(1/\delta)}.
\end{aligned}$$

The proof is completed by choosing $\delta = n^{-2}$ and the fact that $\mathbb{P}(E^c) \leq nK\delta$. □

Frequentist analysis

Bounding the frequentist regret of Thompson sampling is significantly more technical than the Bayesian regret. The trouble is the frequentist regret does not have an expectation with respect to the prior, which means that A_t is not conditionally distributed in the same way as the optimal action (which is not random). For brevity we restrict ourselves to the Gaussian case, but other noise models have also been studied as we discuss at the end of the chapter. To make things simple we assume that $A_t = t$ for $t \in [K]$ and subsequently

$$A_t = \mathrm{argmax}_{i \in [K]}\, \theta_i(t), \tag{35.3}$$

where $\theta_i(t) \sim \mathcal{N}(\hat{\mu}_i(t-1), 1/T_i(t-1))$. Except for the minor detail that we force the algorithm to choose each arm once in the beginning, this policy is derived by

taking an independent Gaussian prior for the mean of each arm and sending the prior variance to infinity.

THEOREM 35.2 *If the algorithm described in Eq. (35.3) is run on Gaussian bandit $\nu \in \mathcal{E}_{\mathcal{N}}^K(1)$, then*

$$R_n \leq C \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{\log(n)}{\Delta_i}\right),$$

where $C > 0$ is a universal constant. Furthermore, $\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}$.

Proof of Theorem 35.2 Recall the notation from Part II that $\hat{\mu}_{1s}$ is the empirical reward of the first arm after s plays of this arm. As usual we assume without loss of generality that $\mu_1 = \max_i \mu_i$ so that the first arm is optimal. Define $Q_s(\varepsilon) = \mathbb{P}\left(\theta_1(t) \geq \mu_1 - \varepsilon \mid T_1(t-1) = s\right)$, which is

$$Q_s(\varepsilon) = \mathbb{P}_{\eta\sim\mathcal{N}(0,1/s)}\left(\hat{\mu}_{1s} + \eta \geq \mu_1 - \varepsilon\right).$$

Let $\varepsilon_1, \ldots, \varepsilon_K$ be a sequence of nonnegative constants to be chosen later and define the event $E_i(t) = \{\theta_i(t) \leq \mu_1 - \varepsilon_i\}$. The plan is to bound $\mathbb{E}[T_i(n)]$ for each suboptimal arm i and then apply Lemma 4.2. We start with a straightforward decomposition.

$$\begin{aligned}\mathbb{E}[T_i(n)] &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{A_t = i\right\}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{A_t = i, E_i(t)\right\}\right] + \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{A_t = i, E_i^c(t)\right\}\right]. \qquad (35.4)\end{aligned}$$

The second sum on the right-hand side is the easy term. Essentially, if $T_i(t-1)$ is large enough, then the probability of $E_i^c(t)$ is unlikely to be very large. We leave it to the reader in Exercise 35.3 to prove for some universal constant $C > 0$ that

$$\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{A_t = i, E_i^c(t)\right\}\right] \leq C\left(1 + \frac{1}{(\Delta_i - \varepsilon_i)^2}\right). \qquad (35.5)$$

The next step is the novel part of the analysis, which bounds the conditional probability that suboptimal arm i is played in round t in terms of the probability of playing the optimal arm. Let $A_t' = \operatorname{argmax}_{i\neq 1} \theta_i(t)$. Then for any $i > 1$,

$$\begin{aligned}\mathbb{P}\left(A_t = 1, E_i(t) \mid \mathcal{F}_{t-1}\right) &\geq \mathbb{P}\left(A_t' = i, E_i(t), \theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1}\right) \\ &= \mathbb{P}\left(\theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1}\right)\mathbb{P}\left(A_t' = i, E_i(t) \mid \mathcal{F}_{t-1}\right) \\ &\geq \frac{Q_{T_1(t-1)}(\varepsilon_i)}{1 - Q_{T_1(t-1)}(\varepsilon_i)}\mathbb{P}\left(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}\right), \qquad (35.6)\end{aligned}$$

where in the first equality we used the fact that $\theta_1(t)$ is conditionally independent

of A'_t and $E_i(t)$ given $\mathcal{F}_{t-1}$. In the second inequality we used the definition of $Q_{T_1(t-1)}(\varepsilon_i) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1})$ and the fact that

$$\mathbb{P}(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}) \leq (1 - \mathbb{P}(\theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1}))\mathbb{P}(A'_t = i, E_i(t) \mid \mathcal{F}_{t-1}) ,$$

which is true because $\{A_t = i, E_i(t)\} \subseteq \{A'_t = i, E_i(t)\} \cap \{\theta_1(t) \leq \mu_1 - \varepsilon_i\}$ and the two intersected events are conditionally independent given $\mathcal{F}_{t-1}$. Therefore using Eq. (35.6) we have

$$\begin{aligned}\mathbb{P}(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}) &\leq \left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1\right)\mathbb{P}(A_t = 1, E_i(t) \mid \mathcal{F}_{t-1}) \\ &\leq \left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1\right)\mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1}) .\end{aligned}$$

Substituting this into the first term in Eq. (35.4) leads to

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}\{A_t = i, E_i(t)\}\right] &\leq \mathbb{E}\left[n \wedge \sum_{t=1}^{n}\left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1\right)\mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1})\right] \\ &= \mathbb{E}\left[n \wedge \sum_{t=1}^{n}\left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1\right)\mathbb{I}\{A_t = 1\}\right] \\ &\leq \mathbb{E}\left[n \wedge \sum_{s=1}^{n}\left(\frac{1}{Q_s(\varepsilon_i)} - 1\right)\right] \\ &\leq \mathbb{E}\left[\sum_{s=1}^{n}\left(n \wedge \left(\frac{1}{Q_s(\varepsilon_i)} - 1\right)\right)\right] . \qquad (35.7)\end{aligned}$$

where in the second last step we used the fact that $T_1(t-1) = s$ is only possible for one round where $A_t = 1$. At last we have decoupled all the dependencies between the arms and reduced the problem to studying the right-hand side of Eq. (35.7). We will shortly show that for any $\gamma \in (0, 1)$,

$$\sum_{s=1}^{n}\mathbb{E}\left[n \wedge \left(\frac{1}{Q_s(\varepsilon_i)} - 1\right)\right] \leq \frac{8\log\left(e + \frac{8}{\varepsilon_i\gamma^2}\right)}{\varepsilon_i^2\gamma^2} + \frac{2\log(n+1)}{\varepsilon_i^2(1-\gamma)} . \qquad (35.8)$$

The theorem follows from the claim and the standard regret decomposition (Lemma 4.2) and by choosing $\varepsilon_i = (1-\gamma)\Delta_i$, where the finite-time result follows with $\gamma = 1/2$ and the asymptotic result with $\gamma = \log^{-1/8}(n)$. The proof of the claim in Eq. (35.8) is a bit of a slog. Let

$$F_s(x) = \sqrt{\frac{s}{2\pi}}\int_{-\infty}^{x}\exp(-sy^2/2)dy$$

be the cumulative distribution function for a Gaussian with zero mean and

variance $1/s$. Then $Q_s(\varepsilon) = 1 - F_s(\mu_1 - \hat{\mu}_{1s} - \varepsilon)$. Therefore

$$\begin{aligned}
\mathbb{E}\left[n \wedge \left(\frac{1}{Q_s(\varepsilon)} - 1\right)\right] &= \int_0^n \mathbb{P}\left(\frac{1}{Q_s(\varepsilon)} - 1 \geq x\right) dx \\
&= \int_0^n \mathbb{P}\left(F_s(\mu_1 - \hat{\mu}_{1s} - \varepsilon) \geq \frac{x}{1+x}\right) dx \\
&= \int_0^n \mathbb{P}\left(\mu_1 - \hat{\mu}_{1s} - \varepsilon \geq F_s^{-1}(x/(1+x))\right) dx \\
&= \int_0^n (1 - F_s(\varepsilon + F_s^{-1}(x/(1+x)))dx\,,
\end{aligned}$$

where in the last line we used the fact that $\mu_1 - \hat{\mu}_{1s}$ is Gaussian with zero mean and variance $1/s$. By Theorem 5.1 it holds that $F_s(-\varepsilon\gamma/2) \leq \exp(-s\varepsilon^2\gamma^2/8)$. Therefore if $x/(1+x) \geq u = \exp(-s\varepsilon^2\gamma^2/8)$, then $F_s^{-1}(x/(1+x)) \geq -\varepsilon\gamma/2$. Abbreviating $g_s(x) = 1 - F_s(\varepsilon + F_s^{-1}(x/(1+x)))$ we see that for $x \geq u/(1-u)$,

$$\begin{aligned}
g_s(x) = \int_x^\infty g_s'(y)dy &= \int_x^\infty \frac{\exp\left(-\frac{s\varepsilon^2 + 2s\varepsilon F_s^{-1}(y/(1+y))}{2}\right)}{(1+y)^2} dy \\
&\leq \int_x^\infty \frac{\exp\left(-\frac{s\varepsilon^2 + 2s\varepsilon F_s^{-1}(x/(1+x))}{2}\right)}{(1+y)^2} dy \leq \int_x^\infty \frac{\exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right)}{(1+y)^2} dy \\
&= \frac{\exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right)}{1+x}\,.
\end{aligned}$$

From its definition it is easily seen that $g_s(x) \leq 1 - x/(1+x) = 1/(1+x)$ so that by splitting the integral we have

$$\begin{aligned}
\mathbb{E}\left[n \wedge \left(\frac{1}{Q_s(\varepsilon)} - 1\right)\right] &= \int_0^n g_s(x)dx \\
&\leq \int_0^{u/(1-u)} \frac{dx}{1+x} + \exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right) \int_{x_0}^n \frac{dx}{1+x} \\
&\leq \log\left(\frac{1}{1 - \exp\left(-\frac{s\gamma^2\varepsilon^2}{8}\right)}\right) + \exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right) \log(n+1)\,. \qquad (35.9)
\end{aligned}$$

We make use of the following facts:

$$\sum_{s=1}^\infty \exp(-sp) \leq \frac{1}{p} \quad \text{and} \quad \sum_{s=1}^\infty \log\left(\frac{1}{1-\exp(-sp)}\right) \leq \frac{\log(e + 1/p)}{p}\,.$$

Summing Eq. (35.9) over s and applying the facts yields the proof of Eq. (35.8):

$$\sum_{s=1}^n \mathbb{E}\left[n \wedge \left(\frac{1}{Q_s(\varepsilon_i)} - 1\right)\right] \leq \frac{8\log\left(e + \frac{8}{\varepsilon_i^2\gamma^2}\right)}{\varepsilon_i^2\gamma^2} + \frac{2\log(n+1)}{\varepsilon_i^2(1-\gamma)}\,. \qquad \square$$

35.2 Linear bandits

While the advantages of Thompson sampling in finite-armed bandits are relatively limited, in the linear setting there is much to be gained, both in terms of computation and empirical performance. Let $\mathcal{E}$ be the set of Gaussian linear bandits with a fixed action-set $\mathcal{A} \subset \mathbb{R}^d$. A Gaussian linear bandit is characterized by its mean vector $\theta \in \mathbb{R}^d$ and the reward after taking action A_t in round t is

$$X_t = \langle A_t, \theta \rangle + \eta_t \,,$$

where $\eta_1, \ldots, \eta_n$ is a sequence of independent standard Gaussian random variables. A prior corresponds to choosing a measure on $\mathbb{R}^d$. An advantage of Thompson sampling relative to optimistic linear bandit algorithms is that the optimization problem for selecting the action no longer requires optimizing over a confidence ellipsoid. There are many cases where this makes a significant difference. For example, if $\mathcal{A}$ is convex, then Thompson sampling can often be computed efficiently, which is not generally the case for the optimistic linear bandit algorithms in Chapter 19.

Input Prior $\mathbb{Q}$ and action-set $\mathcal{A}$
for $t \in 1, \ldots, n$ **do**
    Sample θ_t from the posterior
    Choose $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta_t \rangle$
end for

Algorithm 21: Thompson sampling for linear bandits

The Bayesian regret is controlled using the techniques from the previous section in combination with the concentration analysis in Chapter 20. A frequentist analysis is also possible under slightly unsatisfying assumptions, which we discuss in the notes and bibliographic remarks.

THEOREM 35.3 *Assume that* $\|\theta\|_2 \leq S$ *with* $\mathbb{Q}$*-probability one and* $\sup_{a \in \mathcal{A}} \|a\|_2 \leq L$ *and* $\sup_{a \in \mathcal{A}} \langle a, \theta \rangle \leq 1$ *with* $\mathbb{Q}$*-probability one. Then the Bayesian regret of Algorithm 21 is bounded by*

$$\mathrm{BR}_n \leq 2 + 2\sqrt{2dn\beta^2 \log\left(1 + \frac{nS^2L^2}{d}\right)} \,,$$

where $\beta = 1 + \sqrt{2\log(n) + d\log\left(1 + \frac{nS^2L^2}{d}\right)}$.

Proof We apply the same technique as the proof of Theorem 35.1. Define upper confidence bound $U_t : \mathcal{A} \to \mathbb{R}$ by

$$U_t(a) = \langle \hat{\theta}_{t-1}, a \rangle + \beta \|a\|_{V_{t-1}^{-1}} \,, \quad \text{where } V_t = \frac{I}{\mathbb{E}[\|\theta\|_2^2]} + \sum_{s=1}^{t} A_s A_s^\top \,.$$

By Theorem 20.2, $\mathbb{P}(\text{exists } t \le n : \|\hat{\theta} - \theta\|_{V_t} \ge \beta) \le 1/n$. Let E_t be the event that $\|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}} < \beta$ and $E = \bigcap_{t=1}^n E_t$ and $A^* = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta \rangle$, which is a random variable in this setting because θ is random. Then

$$\begin{aligned}
\mathrm{BR}_n &= \mathbb{E}\left[\sum_{t=1}^n \langle A^* - A_t, \theta \rangle\right] \\
&= \mathbb{E}\left[\mathbb{I}_{E^c} \sum_{t=1}^n \langle A^* - A_t, \theta \rangle\right] + \mathbb{E}\left[\mathbb{I}_E \sum_{t=1}^n \langle A^* - A_t, \theta \rangle\right] \\
&\le 2 + \mathbb{E}\left[\mathbb{I}_E \sum_{t=1}^n \langle A^* - A_t, \theta \rangle\right] \\
&\le 2 + \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}_{E_t} \langle A^* - A_t, \theta \rangle\right] . \qquad (35.10)
\end{aligned}$$

As before, $\mathbb{P}(A^* = \cdot \mid \mathcal{F}_{t-1}) = \mathbb{P}(A_t = \cdot \mid \mathcal{F}_{t-1})$, which means the second term in the above display is bounded by

$$\begin{aligned}
\mathbb{E}_{t-1}\left[\mathbb{I}_{E_t} \langle A^* - A_t, \theta \rangle\right] &= \mathbb{I}_{E_t} \mathbb{E}_{t-1}\left[\langle A^*, \theta \rangle - U_t(A^*) + U_t(A_t) - \langle A_t, \theta \rangle\right] \\
&\le \mathbb{I}_{E_t} \mathbb{E}_{t-1}\left[U_t(A^*) - \langle A_t, \theta \rangle\right] \\
&\le \mathbb{I}_{E_t} \mathbb{E}_{t-1}\left[\langle A_t, \hat{\theta}_{t-1} - \theta \rangle\right] + \beta \|A_t\|_{V_t^{-1}} \\
&\le \mathbb{I}_{E_t} \mathbb{E}_{t-1}\left[\|A_t\|_{V_t^{-1}} \|\hat{\theta}_{t-1} - \theta\|_{V_t}\right] + \beta \|A_t\|_{V_t^{-1}} \\
&\le 2\beta \|A_t\|_{V_t^{-1}} .
\end{aligned}$$

Substituting into the second term of Eq. (35.10),

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}_{E_t} \langle A^* - A_t, \theta \rangle\right] &\le 2\mathbb{E}\left[\beta \sum_{t=1}^n (1 \wedge \|A_t\|_{V_t^{-1}})\right] \\
&\le 2\sqrt{n\mathbb{E}\left[\beta^2 \sum_{t=1}^n (1 \wedge \|A_t\|^2_{V_t^{-1}})\right]} \qquad \text{(Cauchy-Schwartz)} \\
&\le 2\sqrt{2dn\mathbb{E}\left[\beta^2 \log\left(1 + \frac{nS^2L^2}{d}\right)\right]} . \qquad \text{(Lemma 19.1)}
\end{aligned}$$

Putting together the pieces shows that

$$\mathrm{BR}_n \le 2 + 2\sqrt{2dn\beta^2 \log\left(1 + \frac{nS^2L^2}{d}\right)} . \qquad \square$$

Computation

An implementation of Thompson sampling for linear bandits needs to (a) sample θ_t from the posterior and (b) find the optimal action for the sampled parameter:

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta_t \rangle .$$

For some priors and noise models sampling from the posterior is straightforward. The most notable case is when $\mathbb{Q}$ is a multivariate Gaussian. More generally there is a large literature devoted to numerical methods for sampling from posterior distributions. Have sampled θ_t, the optimization problem of finding A_t is a linear optimization problem. Compare this to LinUCB, which needs to solve

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} \operatorname{argmax}_{\tilde{\theta} \in \mathcal{C}} \langle a, \tilde{\theta} \rangle \,,$$

which for large or continuous action-sets is usually much harder computationally.

35.3 Information theoretic analysis

The analysis in the previous sections mirrored those for the frequentist algorithms in Part II. Here we showcase a different approach that relies exclusively on information theory. The argument is based on the observation that for bandits with K arms at most $\log(K)$ nats are needed to code the identity of the optimal arm, which means the total information gain about this quantity is bounded. For many policies one can prove a relationship between the information gain about the optimal arm and the expected regret, which in combination with the previous observation leads to a bound on the regret. This analysis is all the more striking because the assumption that the bandits are (stationary) stochastic can be relaxed as we discuss in the notes.

A few more definitions from information theory are needed. Let X be a discrete random variable on probability space $(\Omega, \mathcal{G}, \mathbb{P})$. Recall from Chapter 14 that the entropy of X is defined by

$$H(X) = \sum_{x \in \operatorname{range}(X)} \mathbb{P}\left(X = x\right) \log \left(\frac{1}{\mathbb{P}\left(X = x\right)} \right) .$$

We also need the **conditional entropy**. Let $\mathcal{F} \subset \mathcal{G}$ be a σ-algebra. Then

$$H(X \mid \mathcal{F}) = \mathbb{E}\left[\sum_{x \in \operatorname{range}(X)} \mathbb{P}\left(X = x \mid \mathcal{F}\right) \log \left(\frac{1}{\mathbb{P}\left(X = x \mid \mathcal{F}\right)} \right) \right] .$$

A little confusingly, the conditional entropy is *not* a random variable. Perhaps a better nomenclature would have been the expected conditional entropy. The entropy of random variable X is a measure of the amount of information in X while the conditional entropy given $\mathcal{F}$ is the expected amount of information required to encode X having observed the information in $\mathcal{F}$. The **mutual information** between X and $\mathcal{F}$ is the difference between the entropy and the conditional entropy:

$$I(X; \mathcal{F}) = H(X) - H(X \mid \mathcal{F}) \,.$$

The mutual information is always nonnegative, which should not be surprising because the information remaining in X can only decrease as more information

is observed. Another name for the mutual information is **information gain**. We use these forms when the underlying measure is clear from context. If this is not the case, then the measure is shown in the subscript: $H(X) = H_{\mathbb{P}}(X)$. The following lemmas provide a chain rule for the mutual information and a simple connection to the relative entropy. The proofs are definitional and are left as exercises.

LEMMA 35.1 *Let X be a random variable on $(\Omega, \mathcal{G}, \mathbb{P})$ and $(\mathcal{F}_t)_{t=0}^n$ a filtration of $\mathcal{G}$ with $\mathcal{F}_0 = \{\emptyset, \mathcal{F}\}$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$. Then*

$$\mathbb{E}\left[\sum_{t=1}^{n} I_{\mathbb{P}_{t-1}}(X; \mathcal{F}_t)\right] = I_{\mathbb{P}}(X; \mathcal{F}_n)\,.$$

LEMMA 35.2 *Let X and Y be random variables on probability space $(\Omega, \mathcal{G}, \mathbb{P})$. If X is discrete and $I(X;Y)$ exists, then*

$$I(X;Y) = \mathbb{E}\left[\mathrm{D}(\mathbb{P}_{Y|X}, \mathbb{P}_Y)\right]\,,$$

where $\mathbb{P}_{Y|X}$ is the random measure on $(\Omega, \sigma(Y))$ such that $\mathbb{P}_{Y|X}(A) = \mathbb{P}(Y \in A \mid X)$ almost surely.

We now present an elegant result connecting the Bayesian regret of any policy and the information gain. Recall that $\mathcal{F}_t = \sigma(A_1, X_{1A_1}, \ldots, A_t, X_{tA_t})$ and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$ and abbreviate $H_t(\cdot) = H_{\mathbb{P}_t}(\cdot)$ and $I_t(\cdot;\cdot) = I_{\mathbb{P}_t}(\cdot;\cdot)$. Define random variable Γ_t to be the ratio of the squared expected Bayesian instantaneous regret and the information gain about the optimal arm.

$$\Gamma_t = \frac{(\mathbb{E}_{t-1}[X_{tA^*} - X_{tA_t}])^2}{I_{t-1}(A^*; (A_t, X_{tA_t}))}\,. \tag{35.11}$$

THEOREM 35.4 *Suppose that $\Gamma_t \le \bar{\Gamma}$ almost surely for all $t \in [n]$. Then*

$$\mathrm{BR}_n \le \sqrt{n\bar{\Gamma}H(A^*)}\,.$$

Proof By the definitions of the regret and Γ_t in Eq. (35.11) and Cauchy-Schwartz we have

$$\begin{aligned}
\mathrm{BR}_n &= \mathbb{E}\left[\sum_{t=1}^{n}(X_{tA^*} - X_{tA_t})\right] = \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_{t-1}[X_{tA^*} - X_{tA_t}]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sqrt{I_{t-1}(A^*; (A_t, X_{tA_t}))\Gamma_t}\right] \le \sqrt{\bar{\Gamma}}\mathbb{E}\left[\sum_{t=1}^{n}\sqrt{I_{t-1}(A^*; (A_t, X_{tA_t}))}\right] \\
&\le \sqrt{n\bar{\Gamma}\mathbb{E}\left[\sum_{t=1}^{n} I_{t-1}(A^*; (A_t, X_{tA_t}))\right]} \le \sqrt{n\bar{\Gamma}H(A^*)}\,,
\end{aligned}$$

where the last inequality follows from Lemma 35.1. □

Theorem 35.4 holds for any policy and clearly illustrates the 'learn something' or 'suffer no regret' argument that appeared in the analyses of so many algorithms. If the ratio of regret relative to information is large, then a policy could suffer high regret. By contrast, policies for which the regret-information ratio is small will enjoy strong regret guarantees.

A combination of Theorem 35.4 and an almost sure bound on Γ_t can lead to nearly optimal bounds on the Bayesian regret of Thompson sampling for finite-armed and linear bandits. We present only the finite-armed case.

LEMMA 35.3 *If $X_{ti} \in [0,1]$ almost surely for all $t \in [n]$ and $i \in [K]$ and A_t is chosen by Thompson sampling using any prior, then $\Gamma_t \leq \frac{K}{2}$ almost surely.*

Before the proof of the lemma we note the consequences. Let $\mathcal{E}$ be a set of finite-armed bandits with K arms and rewards in $[0,1]$. Then Thompson sampling with any prior has its Bayesian regret bounded by

$$\mathrm{BR}_n \leq \sqrt{\frac{Kn \log(K)}{2}}\,. \tag{35.12}$$

Proof of Lemma 35.3 To avoid clutter we drop all subscripts on t. By the chain rule for mutual information we have

$$I(A^*; X_A, A) = I(A^*; A) + I(A^*; X_A \mid A) \tag{35.13}$$

The first term in the above display vanishes because A and A^* are independent under $\mathbb{P}$ (Exercise 35.5).

$$\begin{aligned}
I(A^*; X_A \mid A) &= \sum_{i=1}^{K} \mathbb{P}(A = i)\, I(A^*; X_i) \\
&= \sum_{i=1}^{K} \mathbb{P}(A = i) \sum_{j=1}^{K} \mathbb{P}(A^* = j)\, \mathrm{D}(\mathbb{P}_{X_i \mid A^* = j}, \mathbb{P}_{X_i}) \\
&\geq 2 \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{P}(A = i)\, \mathbb{P}(A^* = j) \left(\mathbb{E}[X_i \mid A^* = j] - \mathbb{E}[X_i]\right)^2 \\
&\geq 2 \sum_{i=1}^{K} \mathbb{P}(A = i)^2 \left(\mathbb{E}[X_i \mid A^* = i] - \mathbb{E}[X_i]\right)^2 \\
&\geq \frac{2}{K} \left(\sum_{i=1}^{K} \mathbb{P}(A = i) \left(\mathbb{E}[X_i \mid A^* = i] - \mathbb{E}[X_i]\right) \right)^2 ,
\end{aligned}$$

where the first equality follows by Eq. (35.13) and the second by Lemma 35.2. The first inequality follows from Pinsker's inequality (Eq. 14.9), the result in Exercise 14.1 and the assumption that the rewards lie in $[0,1]$. The second inequality follows by dropping cross terms and the third by Cauchy-Schwartz. The result follows by rearranging the above display. □

35.4 Notes

1 There are several equivalent ways to view Thompson sampling: *(a)* select an arm according to the posterior probability that it is optimal, *(b)* sample an environment from the posterior and play the optimal action in that environment and *(c)* sample the mean reward for each arm and choose the arm with the largest mean. The algorithms in this chapter is based on *(b)*, but all are equivalent and simply correspond to sampling from different pushforward measures of the posterior. There are three reasons to bear these alternative forms in mind. Historically it seems that Thompson [1933] had the form in *(a)* in mind. The second is computational. Though we are not aware of an example, in some instances beyond finite-armed bandits it may be preferable to sample from a pushforward of the posterior than the posterior itself. The final reason is that in more complicated situations like reinforcement learning it may be desirable to 'approximate' Thompson sampling and approximating a sample from each of the above three choices may lead to different algorithms.

2 Thompson sampling is known to be asymptotically optimal in a variety of settings. Most notably when the noise model follows a single-parameter exponential family and the prior is chosen appropriately [Kaufmann et al., 2012b, Korda et al., 2013]. Unfortunately Thompson sampling is not a golden bullet. The linear variant in Section 35.2 is not asymptotically optimal by the same argument we presented for optimism in Chapter 25. Characterizing the conditions under which Thompson sampling is close to optimal remains an open challenge.

3 For the Gaussian noise model it is known that Thompson sampling is not minimax optimal. Its worst case regret is $R_n = \Theta(\sqrt{nK \log(K)})$ [Agrawal and Goyal, 2013a].

4 An alternative to sampling from the posterior is to choose in each round the arm that maximizes a **Bayesian upper confidence bound**, which is a quantile of the posterior. The resulting algorithm is called **BayesUCB** and has excellent empirical and theoretical guarantees [Kaufmann et al., 2012a, Kaufmann, 2018].

5 The prior has a significant effect on the performance of Thompson sampling. In classical Bayesian statistics a poorly chosen prior is quickly washed away by data. This is not true in bandits because if the prior underestimates the quality of an arm, then Thompson sampling may never play that arm with high probability and no data is ever observed. We ask you to explore this situation in Exercise 35.9.

6 An instantiation of Thompson sampling for linear bandits is known to enjoy near-optimal frequentist regret. In each round the algorithm samples

$\theta_t \sim \mathcal{N}(\hat{\theta}_{t-1}, rV_{t-1})$, where $r = \Theta(d)$ is a constant and

$$V_t = I + \sum_{s=1}^{t} A_s A_s^\top \qquad \text{and} \qquad \hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} X_s A_s \,.$$

Then $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle \theta_t, a \rangle$. This corresponds to assuming the noise is Gaussian with variance r and choosing prior $\mathbb{Q} = \mathcal{N}(0, I)$. Provided the rewards are conditionally 1-subgaussian, the frequentist regret of this algorithm is $R_n = \tilde{O}(d^{3/2}\sqrt{n})$, which is worse than LinUCB by a factor of $\sqrt{d}$. The increased regret is caused by the choice of noise model, which assumes the variance is $r = \Theta(d)$ rather than $r = 1$. The reason to do this comes from the analysis, which works by showing the algorithm is 'optimistic' with reasonable probability. It is not known whether or not this is necessary or an artifact of the analysis. Empirical evidence suggests that $r = 1$ leads to improved performance.

7 A more generic view of Thompson sampling is via the idea of perturbations. The **follow the perturbed leader** algorithm chooses in each round the action

$$A_t = \operatorname{argmax}_{i \in [K]} \left(\hat{\mu}_i(t-1) + \eta_{it} \right) ,$$

where $\eta_{1t}, \ldots, \eta_{Kt}$ is a sequence of independent random variables. In many cases Thompson sampling is hard to analyze because the variance of the randomization is not quite sufficient to prove the optimal arm is optimistic with sufficiently large probability. By sacrificing the Bayesian viewpoint one can sometimes derive a similar algorithm for which the analysis is more straightforward.

8 The analysis in Section 35.3 can be generalized to structured settings such as linear bandits [Russo and Van Roy, 2016]. For linear bandits with an infinite action set the entropy of the optimal action may be infinite. The analysis can be corrected in this case by discretizing the action-set and comparing to a near-optimal action. This leads to a tradeoff between the fineness of the discretization and its size. The algorithm does not depend on the discretization. The reader is referred to the recent article by Dong and Van Roy [2018]. The assumption of bounded rewards in Lemma 35.3 can be relaxed to a subgaussian assumption. For details see the paper by Russo and Van Roy [2016].

9 Nowhere in the proofs of Theorem 35.4 and Lemma 35.3 did we use the fact that bandits in $\mathcal{E}$ are stochastic. Let $\mathcal{E} = [0,1]^{nK}$ and $\mathbb{Q}$ be a prior probability measure on $(\mathcal{E}, \mathfrak{B}(\mathcal{E}))$. We view elements of $\nu \in \mathcal{E}$ as oblivious 'adversarial' bandits, which are really just sequences of reward vectors. Let $(X_{ti})_{ti}$ be a sequence of reward vectors sampled from $\mathbb{Q}$. The optimal action in hindsight is

$$A^* = \operatorname{argmax}_{i \in [K]} \sum_{t=1}^{n} X_{ti} \,.$$

The posterior in round t is $\mathbb{Q}_t = \mathbb{Q}(\cdot \mid A_1, X_{1A_1}, \ldots, A_t, X_{tA_t})$. Then in round t

Thompson sampling samples $\nu \sim \mathbb{Q}_{t-1}$ and chooses $A_t = \operatorname{argmax}_{i \in [K]} \sum_{t=1}^{n} \nu_{ti}$. Repeating the analysis in Section 35.3 shows that

$$\mathbb{E}\left[\sum_{t=1}^{n} X_{tA^*} - X_{tA_t}\right] \leq \sqrt{nK \log(K)/2}\,.$$

The calculation even works for non-oblivious adversaries, provided of course that X_{tA_t} only depends on $A_1, X_1 \ldots, A_{t-1}, X_{t-1}$.

$$\mathbb{P} = \mathbb{Q} \otimes \mathbb{P}_\nu$$

10 The previous note highlights a connection between Bayesian regret and the minimax regret in adversarial bandits. Recall an adversarial Bernoulli finite-armed bandit is a matrix (x_{ti}) with $x_{ti} \in \{0, 1\}$ for all $t \in [n]$ and $i \in [K]$. Let $\mathcal{E}$ be the set of all adversarial bandits and Π the set of all randomized policies and $\mathcal{Q}$ be the set of all distributions on $\mathcal{E}$. Then by the minimax theorem of Sion [1958],

$$\begin{aligned} R_n^* &= \min_{\pi \in \Pi} \max_{(x_{ti}) \in \mathcal{E}} \mathbb{E}_{\pi,x}\left[\max_{i \in [K]} \sum_{t=1}^{n} (x_{ti} - x_{tA_t})\right] \\ &= \max_{Q \in \mathcal{Q}} \min_{\pi \in \Pi} \underbrace{\mathbb{E}_{x \sim Q}\left[\mathbb{E}_{\pi,x}\left[\sum_{t=1}^{n} (x_{ti} - x_{tA_t})\right]\right]}_{\text{Bayesian regret}}\,. \end{aligned}$$

The consequence is that if the regret of the Bayesian optimal algorithm is bounded by B for all priors Q, then the minimax adversarial regret is bounded by B. By Eq. (35.12) we can conclude there exists an adversarial bandit algorithm with worst case regret at most $\sqrt{Kn \log(K)/2}$, which can be strengthened using the result in Exercise 35.2. Of course we already knew these things, but the approach has applications in more sophisticated settings. The most notable example being the first near-optimal analysis for adversarial convex bandits [Bubeck et al., 2015a, Bubeck and Eldan, 2016]. The main disadvantage is that uniform bounds on the Bayesian regret implies existence of a single algorithm with small minimax adversarial regret, but the result is nonconstructive.

11 The information-theoretic ideas in Section 35.3 suggest that rather than sampling A_t from the posterior on A^*, one can sample A_t from the distribution minimizing Eq. (35.11). Specifically, A_t is sampled from distribution π_t on $[K]$ where

$$\pi_t = \operatorname{argmin}_\pi \frac{\left(\sum_{i=1}^{K} \pi(i)(\mathbb{E}_{t-1}[X_{tA^*} - X_{ti}])\right)^2}{\sum_{i=1}^{K} \pi(i) I_{t-1}(A^*; X_{ti} \mid A_t = i)}\,.$$

The resulting policy is called **Information Directed Sampling**. Bayesian regret analysis for this algorithm follows along similar lines as what was presented in Section 35.3. See the paper by Russo and Van Roy [2014a] for more details or Exercise 35.7.

35.5 Bibliographic remarks

Thompson sampling has the honor of being the first bandit algorithm and is named after its inventor [Thompson, 1933], who considered the Bernoulli case with two arms. Thompson provided no theoretical guarantees, but argued intuitively and gave hand-calculated empirical analysis. It would be wrong to say that Thompson sampling was entirely ignored, but its popularity soared when a large number of authors independently rediscovered the article/algorithm [Granmo, 2010, Ortega and Braun, 2010, Graepel et al., 2010, Chapelle and Li, 2011, May et al., 2012]. The surge in interest was mostly empirical, but theoreticians followed soon with regret guarantees. For the frequentist analysis we followed the proofs by Agrawal and Goyal [2013a, 2012], but the setting is slightly different. We presented results for the 'realizable' case where the payoff distributions are actually Gaussian, while Agrawal and Goyal use the same algorithm but prove bounds for rewards bounded in $[0, 1]$. Agrawal and Goyal [2013a] also analyze the Beta/Bernoulli variant of Thompson sampling, which for rewards in $[0, 1]$ is asymptotically optimal in the same way as KL-UCB (see Chapter 10). This result was simultaneously obtained by Kaufmann et al. [2012b], who later showed that for appropriate priors asymptotic optimality holds for single parameter exponential families [Korda et al., 2013]. For Gaussian bandits with unknown mean and variance Thompson sampling is asymptotically optimal for some priors, but not others – even quite natural ones [Honda and Takemura, 2014]. The Bayesian analysis of Thompson sampling based on confidence intervals is due to Russo and Van Roy [2014b] while the information-theoretic argument is by Russo and Van Roy [2014a, 2016]. Recently the idea has been applied to a wide range of bandit settings [Kawale et al., 2015, Agrawal et al., 2017] and reinforcement learning [Osband et al., 2013, Gopalan and Mannor, 2015, Leike et al., 2016, Kim, 2017]. The BayesUCB algorithm is due to Kaufmann et al. [2012a] with improved analysis and results by Kaufmann [2018]. The frequentist analysis of Thompson sampling for linear bandits is by Agrawal and Goyal [2013b] with refined analysis by Abeille and Lazaric [2017a] and a spectral version by Kocák et al. [2014]. There is a tutorial on Thompson sampling by Russo et al. [2017] that focuses mostly on applications and computational issues.

35.6 Exercises

35.1 Consider the event E defined in Theorem 35.1 and prove that $\mathbb{P}(E^c) \leq nK\delta$.

35.2 Improve the bound in Theorem 35.1 to show that $\mathrm{BR}_n \leq C\sqrt{Kn}$ where $C > 0$ is a universal constant.

Replace the naive confidence intervals used in the proof of Theorem 35.1 by the more refined confidence bounds used in Chapter 9. The source for this result is the paper by Bubeck and Liu [2013].

35.3 Prove the inequality in Eq. (35.5).

35.4 Prove Lemmas 35.1 and 35.2.

35.5 Suppose that X and Y are independent random variables. Show that $I(X;Y) = 0$.

35.6 Let $\mathcal{E}$ be a set of bandits and $\mathbb{Q}$ a prior on $\mathcal{E}$.

(a) Recall that $R_n^*(\mathcal{E}) = \inf_\pi \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$ is the minimax regret. Show that $R_n^*(\mathcal{E}) \geq \inf_\pi \mathrm{BR}_n(\mathcal{E}, \mathbb{Q})$.
(b) Let $\mathcal{E}$ be the set of Bernoulli bandits. Find a sequence of priors $(\mathbb{Q}_n)$ such that $\mathrm{BR}_n(\mathcal{E}, \mathbb{Q}_n) \geq c\sqrt{Kn}$ for all $n \geq K$ where $c > 0$ is a universal constant.

35.7 Prove that for any prior such that $X_{ti} \in [0,1]$ almost surely the Bayesian regret of information-directed sampling satisfies

$$\mathrm{BR}_n \leq \sqrt{\frac{Kn \log(K)}{2}}\,.$$

35.8 The purpose of this exercise is to compare Thompson sampling for Gaussian bandits with UCB.

(a) Implement the Gaussian Thompson sampling algorithm described by Eq. (35.3).
(b) Compare the expected regret of Thompson sampling with the version of UCB in Chapter 8 and refinements in Eq. (9.2) and Eq. (9.3).
(c) What about the variance of these algorithms?
(d) Briefly explain the pros and cons of Thompson sampling relative to UCB.

35.9 Fix a Gaussian bandit with unit variance and mean vector $\mu = (0, 1/10)$ and horizon $n = 1000$. Now consider Thompson sampling with a Gaussian model with known unit covariance and a prior on the unknown mean of each arm given by a Gaussian distribution with mean μ_P and covariance $\sigma_P^2 I$.

(a) Let the prior mean be $\mu_P = (0,0)$ and plot the regret of Thompson sampling as a function of the prior variance σ_P^2.
(b) Repeat the above with $\mu_P = (0, 1/10)$ and $(0, -1/10)$ and $(2/10, 1/10)$.
(c) Explain your results.