EE194/290S: ML for Sequential Decision under Uncertainty **Fall 2018**

Lecture 21: Thompson Sampling; Contextual Bandits

Instructors: Anant Sahai, Vidya Muthukumar

Scribes: Kailas Vodrahalli, Vignesh Subramanian

1 Introduction

Before proceeding, we define the notation we will be using throughout these notes (following the lecture). Please refer to these definitions when studying these notes.

- $\mathcal{A}$ is the set of possible actions. We assume $|\mathcal{A}|$ is finite, though it is not required to be in general.
- A^* is the true optimal action in hindsight (we assume a fixed time horizon throughout).
- A_t is a random variable describing our choice of action at time t.
- a_t is a specific realization of A_t.
- θ is the hidden state of the Bandit system.
- Y_{t,A_t} is a random variable describing our observation at time t. Note the sources of randomness: A_t, θ, and the observation itself.
- Y_{t,a_t} is a realization of Y_{t,A_t}.
- $\mathbf{Y_t}$ is a vector containing the resulting observations from each possible action at time t. If we observe $\mathbf{Y_t}$ at each time step, we are in the "Prediction" setting.

1.1 Recall Multi-armed bandit

First we recall the Multi-armed bandit setting. Figure 1 contains a visualization of the setup. We will revisit and revise this framework when we introduce Contextual bandits at the end of these notes.

1. We select an action $a_t \in \mathcal{A}$, where $|\mathcal{A}|$ is finite.
2. The Bandit system produces an observable output based on our action, Y_{t,a_t}. The Bandit system is parameterized by some hidden state θ; our observation contains information on this hidden state.
3. We compute a reward, $R(Y_{t,a_t})$. From this reward, we can update our action selection method in the next round.

Also note that in general, Y_{t,a_t} can be thought of as coming from $\mathbf{Y_t}$ and contains limited information on θ based on our action.

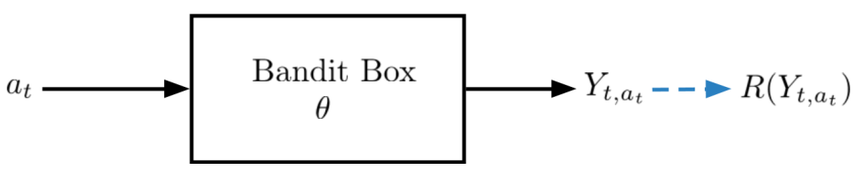


Figure 1: Caption

1.2 Recall Thompson Sampling

Thompson Sampling is a framework for selecting a_t in step (1) and updating our selection method in step (3). We define the random variable A_t such that $P(A_t = a) = P(A^* = a|(A_1, Y_{1,A_1}), \ldots, (A_{t-1}, Y_{t-1,A_{t-1}})), \forall a \in \mathcal{A}$. We then sample A_t to get our action in step (1). In words, we are sampling from the posterior of the optimal action conditioned on our observed past.

Last lecture, we found that a bound on the information ratio gives us a bound on the regret, where the information ratio is defined as

$$\Gamma_t = \frac{(\mathbb{E}_t[R(Y_{A^*}) - R(Y_{A_t})])^2}{I_t(A^*; (A_t, Y_{t,A_t}))} \tag{1}$$

In particular, if there exists a constant $\bar{\Gamma}$ such that $\Gamma_t \leq \bar{\Gamma}$, then

$$\mathbb{E}[\text{Regret}(T)] \leq \sqrt{\bar{\Gamma} \cdot T \cdot H(A^*)} \tag{2}$$

where Regret is defined relative to the optimal action in hindsight and $H(\cdot)$ is the entropy of a discrete random variable. Note here that $H(A^*) = I_0(A^*;)$, the mutual information before we have any observations.

Also recall that last time we showed that in the full feedback case with $Y_{t,a_t} = \mathbf{Y_t}$, $\bar{\Gamma} \leq 2$. We will now proceed to provide a bound in the more interesting limited feedback case we typically have in a multi-armed bandits setting.

2 Proof of Upper Bound for Thompson Sampling

Our goal here will be to show that

$$\Gamma_t \leq 2|\mathcal{A}|. \tag{3}$$

With a more delicate analysis this bound can be constructed with tighter constants , but it turns out the dependence on the action set size will remain.

Aside: When the action space $\mathcal{A}$ is continuous, this bound is not very meaningful in its current form since $H(A^*)$ is infinity. One workaround is to take a quantized approximation approach and discretize the set $\mathcal{A}^*$ for the purposes of analysis.

2.1 Information Ratio Bound

Note: For brevity of notation, in the proof we suppress the subscript t and denote Y_{t,A_t} as Y_{A_t}. We start with the information ratio as defined in equation 1 and expand.

$$\Gamma_t = \frac{(\mathbb{E}_t[R(Y_{A^*}) - R(Y_{A_t})])^2}{I_t(A^*; (A_t, Y_{A_t}))} \tag{4}$$

$$\overset{(a)}{=} \frac{(\sum_{a\in\mathcal{A}} P_t(A^* = a) \cdot \mathbb{E}_t[R(Y_a)|A^* = a] - \sum_{a\in\mathcal{A}} P_t(A_t = a) \cdot \mathbb{E}_t[R(Y_a)|A_t = a])^2}{I_t(A^*; (A_t, Y_{A_t}))}$$

$$\overset{(b)}{=} \frac{(\sum_{a\in\mathcal{A}} P_t(A^* = a) \cdot (\mathbb{E}_t[R(Y_a)|A^* = a] - \mathbb{E}_t[R(Y_a)]))^2}{I_t(A^*; (A_t, Y_{A_t}))}. \tag{5}$$

(a) : The expectation $\mathbb{E}_t$ in the definition of Γ_t is over the randomness in A^*, A_t and the observations $\mathbf{Y_t}$. We condition on $A*$ and A_t respectively in the two terms and expand the expectations.

(b) : For Thompson sampling $P_t(A_t = a) = P_t(A^* = a)$. Furthermore, action A_t is selected based on the past observations and actions and thus conditioned on the past, $\mathbf{Y}_t$ and A_t are independent. From here it follows that $R(Y_a)$ and A_t are conditionally independent given the past and $\mathbb{E}_t[R(Y_a)|A_t = a] = \mathbb{E}_t[R(Y_a)]$.

We would like to find an upper bound on the numerator of the form $2|\mathcal{A}| \cdot (I_t(A^*; (A_t, Y_{A_t}))$ so that after cancelling with the term from the denominator we get the desired bound. To do this let us expand the denominator and express it in terms of the KL divergence between two probability distributions.

$$\begin{aligned} I_t(A^*; (A_t, Y_{A_t})) &\overset{(c)}{=} \sum_{a \in \mathcal{A}} P_t(A_t = a) I_t(A^*; Y_a) \\ &\overset{(d)}{=} \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} P_t(A^* = a) \cdot P_t(A^* = a') \cdot 2 \cdot D_{KL}(P_t(Y_a | A^* = a') || P_t(Y_a)). \end{aligned} \tag{6}$$

(c) : A^* is conditionally independent of A_t given the past and thus $I_t(A^*; (A_t, Y_{A_t})) = I_t(A^*; Y_{A_t})$. We then condition $I_t(A^*; Y_{A_t})$ on the action A_t.

(d) : We replace $P_t(A_t = a)$ with $P_t(A^* = a)$ since the two are equal for Thompson sampling and observe that mutual information can be expressed in terms of KL Divergence as,

$$I_t(A^*; Y_a) = \sum_{a' \in \mathcal{A}} P_t(A^* = a') \cdot D_{KL}(P_t(Y_a | A^* = a') || P_t(Y_a)).$$

We continue from equation 5 and observe that to get a bound on the numerator of the form in 6 we have to introduce a double summation over two independent variables. We will do this in two steps. First we invoke the Cauchy Schwartz inequality:

$$(\mathbf{u}^T \mathbf{v})^2 \leq (\mathbf{u}^T \mathbf{u})(\mathbf{v}^T \mathbf{v}),$$

with $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{|\mathcal{A}|}$, $\mathbf{u}[a] = 1$, $\mathbf{v}[a] = P_t(A^* = a)(\mathbb{E}_t[R(Y_a)|A^* = a]\mathbb{E}_t[R(Y_a)])$.
Thus,

$$\begin{aligned} (\mathbb{E}_t[R(Y_{A^*}) - R(Y_{A_t})])^2 &= (\sum_{a \in \mathcal{A}} P_t(A^* = a) \cdot (\mathbb{E}_t[R(Y_a)|A^* = a] - \mathbb{E}_t[R(Y_a)]))^2 \\ &\leq |\mathcal{A}| \cdot (\sum_{a \in \mathcal{A}} P_t(A^* = a)^2 (\mathbb{E}_t[R(Y_a)|A^* = a] - \mathbb{E}_t[R(Y_a)])^2). \end{aligned} \tag{7}$$

Next we can view the summation term in the above equation as the result of summing over the diagonal elements of a two dimensional matrix of size $|\mathcal{A}| \times |\mathcal{A}|$ where one dimension is indexed by a and the other is indexed by a'. We are free to fill up rest of the entries in the matrix as per our choice and as long as each element is non-negative we will preserve the inequality. We construct the matrix to have entries

$$M(a, a') = P_t(A^* = a) \cdot P_t(A^* = a') \cdot (\mathbb{E}_t[R(Y_a)|A^* = a'] - \mathbb{E}_t[R(Y_a)])^2.$$

Note that on the diagonal $a = a'$ and we get back the original terms in the summation in equation 7. Continuing the bound we get,

$$\begin{aligned} &(\mathbb{E}_t[R(Y_{A^*}) - R(Y_{A_t})])^2 \\ &\leq |\mathcal{A}| \cdot (\sum_{a \in \mathcal{A}} P_t(A^* = a)^2 (\mathbb{E}_t[R(Y_a)|A^* = a] - \mathbb{E}_t[R(Y_a)])^2) \\ &\leq |\mathcal{A}| \cdot (\sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} P_t(A^* = a) \cdot P_t(A^* = a') \cdot (\mathbb{E}_t[R(Y_a)|A^* = a'] - \mathbb{E}_t[R(Y_a)])^2) \\ &\overset{(e)}{\leq} |\mathcal{A}| \cdot (\sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} P_t(A^* = a) \cdot P_t(A^* = a') \cdot 2 \cdot D_{KL}(P_t(Y_a | A^* = a') || P_t(Y_a))). \end{aligned} \tag{8}$$

(e) : From last lecture the following bound relating regret to the KL divergence,

$$(\mathbb{E}_t[R(Y_a)|A^* = a'] - R(Y_a)])^2 \leq 2 \cdot D_{KL}(P_t(Y_a | A^* = a') || P_t(Y_a)).$$

Note that we invoke the Cauchy Schwartz inequality first to give us more flexibility in choosing the non-diagonal entries of the matrix in the subsequent step. Putting together equations 4, 6 and 8 we complete the proof and bound the information ratio as,

$$\Gamma_t \leq 2|\mathcal{A}|. \tag{9}$$

2.2 Regret Bound

Thus we have shown that the information ratio is bounded. Using our earlier result, this bound implies that

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] &\leq \sqrt{\bar{\Gamma} \cdot T \cdot H(A^*)} \\ &\leq \sqrt{2|\mathcal{A}| \cdot T \cdot H(A^*)}.\end{aligned}$$

2.3 Computational vs Information Efficiency

Below we include the pseudocode for Thompson sampling for multi-armed bandits. We assume Bernoulli bandits and a beta-distribution prior over each bandit for concreteness to demonstrate computational efficiency (for the more general case, simply adjust how reward and how posterior are computed). We choose a beta prior because updating the posterior is very concise. A beta distribution is parametrized by 2 parameters, so we let θ_t be a $k \times 2$ array in the pseudocode below. Visualizations for this example are included in section 3.

Note that to sample over $P(A^* = a)$, we first sample over parameter distributions of each bandit and then select the bandit that maximizes the expected reward given its sampled distribution.

Algorithm 1 Thompson sampling for multi-armed Bernoulli bandits with beta prior

procedure THOMPSON($s_1, s_2, \ldots, s_k, T$) ▷ $s_i[t]$ is the outcome of bandit i at time t
  $\theta_0[i] \leftarrow$ prior distribution over parameters for bandit i
  for $t = 1, 2, \ldots, T$ **do**
    $\hat{\theta}[i] \leftarrow$ sample from $\theta_{t-1}[i]$
    $action_t \leftarrow \arg\max_i \hat{\theta}[i]$ ▷ select bandit that has highest expected reward given sampled parameters
    $reward_t \leftarrow s_{action_t}[t]$ ▷ reward is sampled bandit's outcome (0/1)
    $\theta_t \leftarrow \theta_{t-1}$
    $\theta_t[action_t] \leftarrow [reward_t, 1 - reward_t]$ ▷ Update posterior for beta distribution

So looking at this algorithm, at each time step we need to (1) sample from a k beta distributions, (2) compute an argmax over k elements, and (3) update our posterior. Assuming sampling from a beta distribution is constant time, the runtime at each iteration is $O(k)$, which is as efficient as we can hope for if we want to consider all k bandits at every iteration.

However, Thompson sampling is **not** optimal in an information sense. Consider that we are still exploring at time T despite knowing we have a fixed time horizon. Intuitively, we can attain a smaller expected regret by just being greedy at the last time step and choosing the bandit whose expected posterior gives the bandit with highest expected reward (i.e., there is no benefit to exploring anymore, so don't).

3 Visualization of Upper Bound

To visualize Thompson sampling, we will consider the problem where we have two Bernoulli bandits with unknown means. We assume beta-distributed priors, as updating the posterior is easy in this case. We visualize empirical regret and pseudoregret. Recall that regret and pseudoregret are defined in this context as

$$Regret = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} Y_{t,a} - \sum_{t=1}^{T} Y_{t,a_t} \tag{10}$$

$$Pseudoregret = \sum_{t=1}^{T} Y_{t,a^*} - \sum_{t=1}^{T} Y_{t,a_t} \tag{11}$$

where a^* is defined as

$$a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[\sum_{t=1}^{T} Y_{t,a}]. \tag{12}$$

The pseudocode is below (we allow for k bandits):

3.1 Heatmap visualization

Here we provide some visualizations for Thompson sampling. We are in the setting with 2 Bernoulli bandits; for each we have a beta distribution prior. For more discussion on beta distribution priors in the Bernoulli bandits setting, we recommend reading through [2], a well-written article containing a theoretical and practical discussion on Thompson sampling.

In the heatmap visualizations below, the X and Y axes each correspond to the true Bernoulli mean for one of the bandits. The color of the heatmap reflects a value described in the caption for each plot. Values are averaged over 20 random initializations.

Note the difference in regret between the full information (prediction) setting and the multi-armed bandit setting. Full information allows for lower regret, as it takes fewer steps to determine which bandit has the higher Bernoulli mean.

Also note where regret is largest in both full information and multi-armed bandit settings. It is at locations where the bandits are nearly equal. At these points, it is difficult to determine which bandit has higher mean, and so we incur higher regret despite expected regret at each time step being smaller. However, pseudoregret along the diagonals is near 0. This is because pseudoregret selects the arm to compare against based on the expected reward rather than empirical values.

In Figure 6, we show this phenomenon more clearly. Here, we again have 2 Bernoulli bandits with beta priors. The X axis reflects the value of ϵ: we set the the mean of bandit 1 to $0.5 - \epsilon$ and the mean of bandit 2 to $0.5 + \epsilon$. We plot total regret for various time horizons. Note that for small ϵ, regret is largest for larger time horizons; for smaller time horizons, the largest regret is incurred at slightly larger values of ϵ. The intuition here is that there is a tradeoff between small ϵ values incurring small regret each iteration, but incurring this regret for exponentially more iterations than larger ϵs. Because of this tradeoff, the ϵ that corresponds to largest regret depends on the specific time horizon.

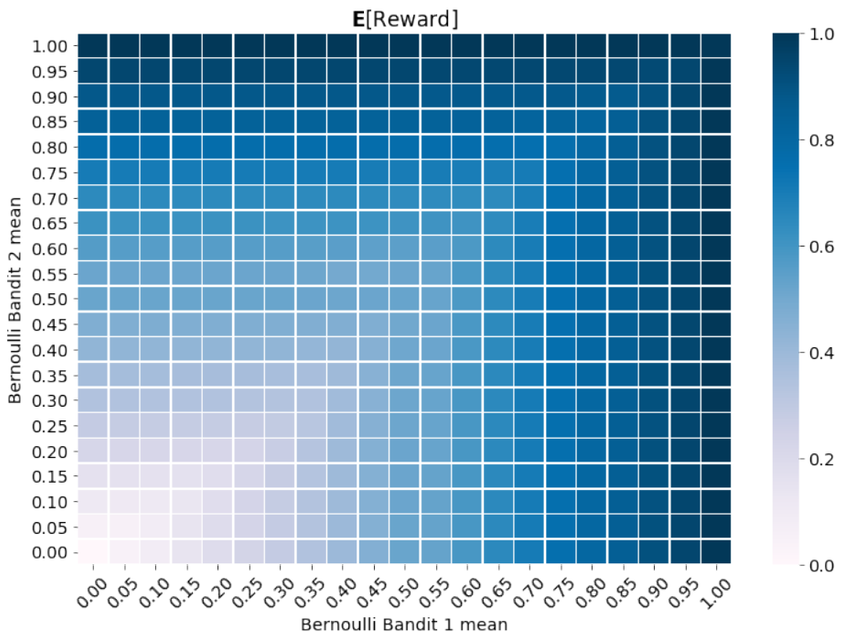
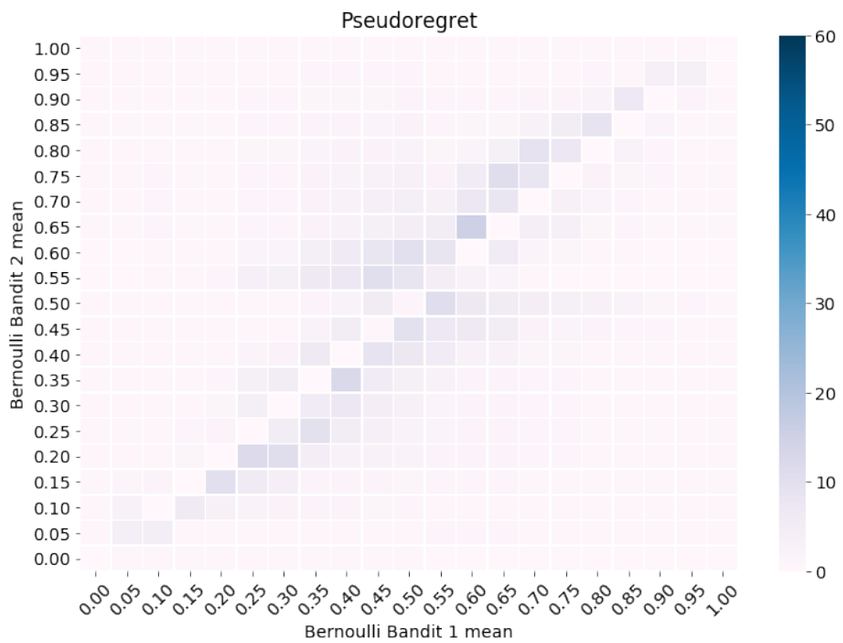


Figure 2: Expected reward using Thompson Sampling in Bernoulli Bandit setting. Use beta distribution prior.

4 Introduction to Contextual Bandits

Next we switch gears and introduce the contextual bandits problem. Contextual bandits is a generalization of the multi-armed bandits setting where we additionally have a time-varying

(a) Regret



(b) Pseudoregret

Figure 3: Regret/Pseudoregret after $T = 1000$ using Thompson Sampling in Bernoulli Bandit setting with full information. Use beta distribution prior. Notice the diagonal is near zero for pseudoregret but not for regret

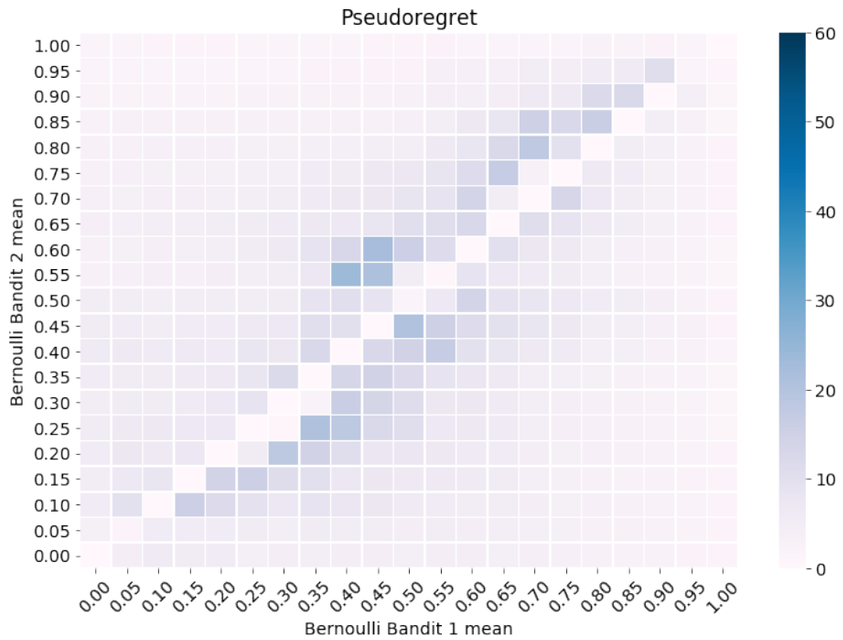
context associated with the bandit system. The optimal action at each time step is no longer the same but instead depends on the context. We will consider a policy π to be a mapping from contexts to actions that determines how we choose to act based on the context we receive.

A real world example of the contextual bandit problem is that of a website like Facebook deciding which ad to display. Here the context is the preferences, likes, and dislikes of the user as well as the page the user is on. Facebook decides the ad to be shown to the user based on this context.

To define regret in the contextual bandit setting we will restrict the best in-hindsight actions to those from a set of reference policies denoted Π. For instance, Π could be the set of all policies based on a decision tree of depth 1.

Regret in the contextual banding setting is defined by comparing the reward of our policy

(a) Regret



(b) Pseudoregret

Figure 4: Regret/Pseudoregret after $T = 1000$ using Thompson Sampling in Bernoulli Bandit setting. Use beta distribution prior. Notice the diagonal is near zero for pseudoregret but not for regret

to that of the best policy in Π as follows,

$$\mathbb{E}[\text{Regret}(T)] = \mathbb{E}[\max_{\pi \in \Pi} \sum_{t=1}^{T} R(X_t, Y_{t,\pi(X_t)}) - \sum_{t=1}^{T} R(X_t, Y_{t,A_t})]. \tag{13}$$

As an aside it is interesting to look at the connection between context and state. State can be viewed as the true parameter describing the underlying environment while context could be a partial or noisy observation of the state. The ideal context is the true state but we rarely have access to this in practical settings.

As we have seen before the online learning setting is an extension to the prediction setting and contextual bandits is an extension to the multi-armed bandits setting. A key difference between the prediction setting and the multi-armed bandits setting was the exploration exploitation trade-off and this persists when we compare contextual bandits to online learning. In the online learning setting once we get the true label, we can compute what our rewards would have been for any prediction that we could have made. However in contextual bandits

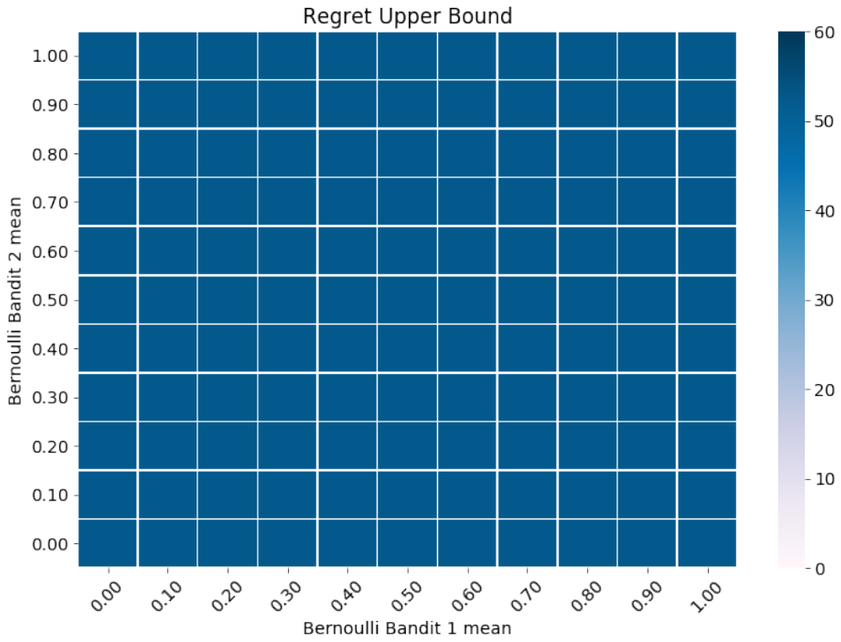


Figure 5: Regret upper bound on Thompson Sampling in Bernoulli Bandit setting.

we only get information about reward for the action we decide to take and no information about rewards for other actions.

Since we know how to solve the multi-armed bandits problem we next explore if we can cast the contextual bandit problem as a version of the multi-armed bandit problem by viewing the policies as arms. We will see that this approach has several drawbacks and naively following this approach may lead to very large regret in the general case. First we observe that if our class of policies in unrestricted then the number of policies grows exponentially in the number of features in our context. Consider the case where our context consists of k binary features, i.e. $X_t \in \{0,1\}^k$. The set of policies Π, contains all mappings $\pi : \{0,1\}^k \mapsto \mathcal{A}$ and its cardinality is given by $|\Pi| = |\mathcal{A}|^{2^k}$. If we naively cast the contextual problem as the multi-armed bandits problem by viewing policies as arms then we would get regret of the form,

$$\mathbb{E}[\text{Regret}(T)] = \sqrt{4 \cdot |\mathcal{A}|^{2^k} \cdot T \cdot \log |\mathcal{A}|^{2^k}}. \tag{14}$$

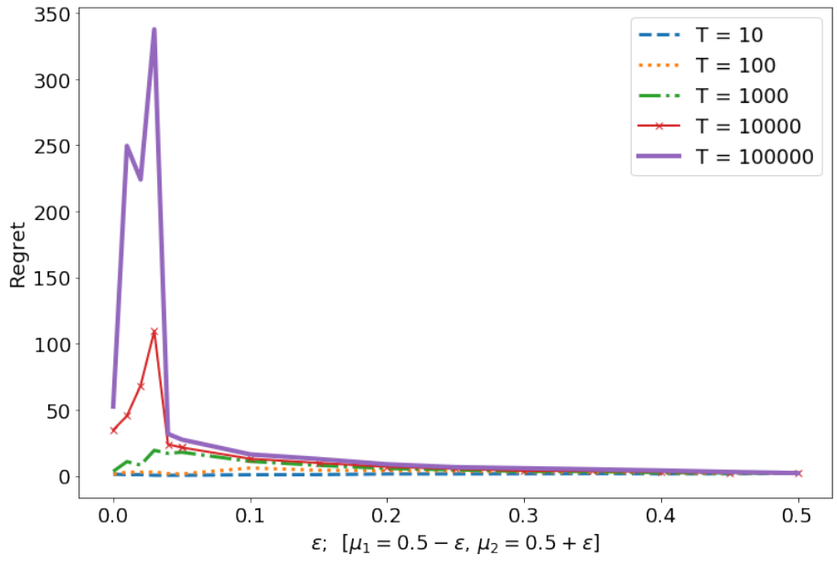
Note that the effective number of actions is now $|\mathcal{A}|^{2^k}$ instead of $|\mathcal{A}|$. The term $|\mathcal{A}|^{2^k}$, even for moderate values of $|A|$ and k, is very large and impractical. We would like to achieve a better bound.

A key point to note here is that policies differ from arms in the sense that taking an action allows us to learn about all policies that would have played this action, and we can learn about multiple policies from each action. This is in contrast to how each action allows us to learn only about one arm in the Bernoulli bandits setting we have studied. In subsequent lectures we will see how we can use this idea to get a regret bound of the form

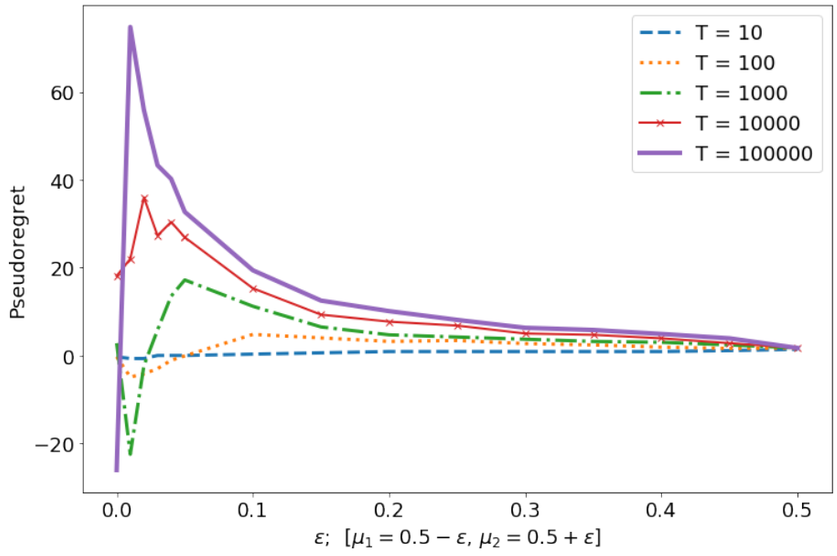
$$\mathbb{E}[\text{Regret}(T)] = \sqrt{4 \cdot |\mathcal{A}| \cdot T \cdot \log |\mathcal{A}|^{2^k}}, \tag{15}$$

where the exponential dependence on the number of context features persists only inside the log term. This means that regret will depend on the square root of 2^k, which represents (in some sense) the complexity of the policies we are willing to allow.

We end here with a final example where we note that Thompson sampling can be used to solve the contextual bandits problem. As an example, consider the following (concrete) scenario: Context 1 occurs with probability 0.4 and context 2 occurs with probability 0.6. In context 1, we have 2 Bernoulli bandits with means $\mu_{11} = 0.3$ and $\mu_{12} = 0.7$. In context 2 we have 2 Bernoulli bandits with means $\mu_{21} = 0.6$ and $\mu_{22} = 0.5$. We consider all 4 possible deterministic policies for computing regret and for selecting our actions. We assume beta priors over policies. Then the following pseudocode describes a Thompson sampling based algorithm to solve this problem.



(a) Regret



(b) Pseudoregret

Figure 6: Plot of regret / pseudoregret when for various time horizons T and varying ϵ. Binary bandits setting where $\mu_1 = 0.5 - \epsilon$ and $\mu_2 = 0.5 + \epsilon$. In the pseudoregret plot, the dip at the start is and artifact of computing the pseudoregret empirically (expected value of pseudoregret when $\epsilon = 0$ is 0).

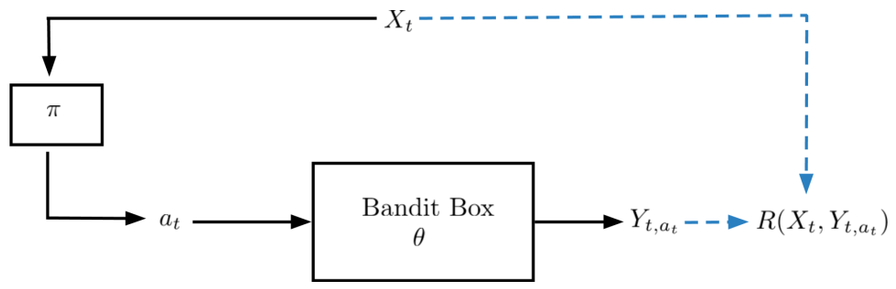


Figure 7: Caption

Note that we consider **all** possible policies: this means we have 2^k policies where k is the number of bandits. Also note that we let our priors be over the policies rather than over actions. This is valid here since we consider all possible policies; in general, we may need to maintain priors over actions.

We plot the described scenario in Figure 8. As can be seen on the curve, we have less than square root regret in T!

Algorithm 2 Thompson sampling for contextual Bernoulli bandits with beta prior

procedure TS-CONTEXT($s_1, s_2, \ldots, s_k, T$)
  $\theta_0[i] \leftarrow$ prior distribution over parameters for policy i
  for $t = 1, 2, \ldots, T$ **do**
    $x_t \leftarrow$ context at time t
    $\hat{\theta}[i] \leftarrow$ sample from $\theta_{t-1}[i]$
    $policy_t \leftarrow \arg\max_i \hat{\theta}[i]$ ▷ select policy that has highest expected reward given sampled parameters
    $action_t \leftarrow policy_t(x_t)$ ▷ obtain action from policy
    $reward_t \leftarrow s_{action_t}[t]$ ▷ reward is sampled bandit's outcome (0/1)
    $\theta_t \leftarrow \theta_{t-1}$
    for $i = 1, \ldots,$ number of policies **do** ▷ Update all policies that took our action
      $action_i \leftarrow policy_i(x_t)$
      if $action_t == action_i$ **then**
        $reward_i \leftarrow s_{action_i}[t]$
        $\theta_t[action_i] \leftarrow [reward_i, 1 - reward_i]$ ▷ Update posterior for beta distribution

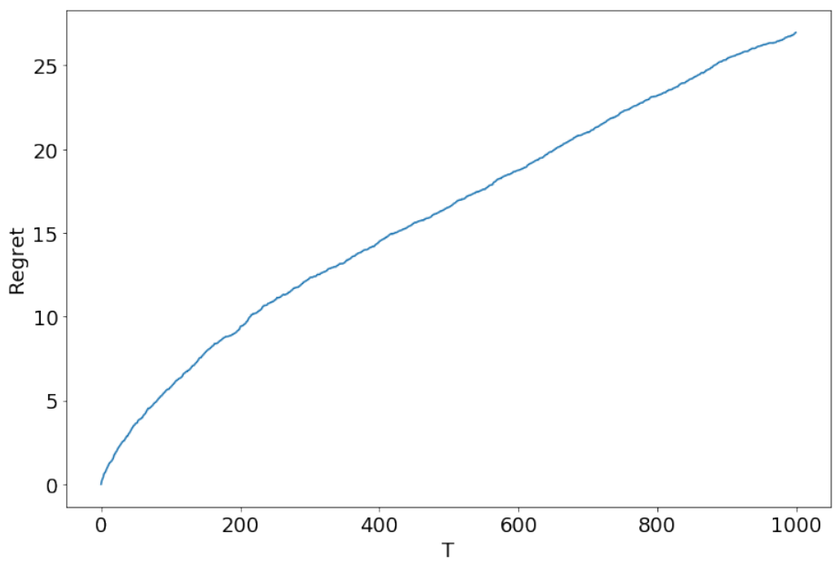


Figure 8: Plot of regret over T (averaged across 100 initializations) in contextual bandits setting.

References

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

[2] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.