# Coding exercise: Bandits
# Foundations of Machine Learning - Oxford Summer School 2025

Maximilian Kasy

In this problem, you are asked to implement some simulations and estimators in R, or in Python. Your code should run from start to end in one execution, producing all the output. Output and discussion of findings should be integrated in a report generated in R-Markdown, or from a Jupyter Notebook. Figures and tables should be clearly labeled and interpretable. The findings should be discussed in the context of the theoretical results that we derived in class.

1. In this problem, you are asked to simulate data for a Bernoulli bandit problem, where

$$D_t \in \{1, \dots, k\}, \qquad Y_t = Y^{D_t}, \qquad Y_t^d \sim Ber(\theta^d).$$

   and treatment is assigned using Thompson sampling with a uniform prior, $(\theta^1, \dots, \theta^k) \sim U([0,1]^k)$. Recall that Thompson sampling assigns

$$D_t = \underset{d}{\operatorname{argmax}}\ \hat{\theta}_t^d,$$

   where $\hat{\theta}_t$ is a draw from the posterior after period $t-1$.

   (a) Set up a function which accepts a sample size $T$ and a $k$-vector $(\theta^1, \dots, \theta^k)$ as its arguments, and returns a history $(D_t, Y_t)_{t=1}^T$ generated based on the Bernoulli bandit model and Thompson sampling.

   (b) Write a second function which takes the same arguments, plus a number of replications $R$, and evaluates the first function $R$ times (using parallel computing; for instance the *future* package).

   This function should return 4 vectors of length $T$: The averages of $Y_t$, $\theta^{D_t}$, $\mathbf{1}(D_t = \operatorname{argmax}_d \theta^d)$, and $\max \theta^d - \theta^{D_t}$, for each time periord $t$.

   (c) Pick a fixed vector of parameters $(\theta^1, \dots, \theta^k)$ and a time horizon $T$ and use the second function to plot the average (across replications) of cumulative average regret

$$\tfrac{1}{T} \sum_{1 \le t \le T} \left[ \left( \max_d \theta^d \right) - \theta^{D_t} \right]$$

   as a function of $T$, using a large number of replications $R$ (such as $R = 10.000$). Repeat this for several different choices of $(\theta^1, \dots, \theta^k)$.

   How does the result relate to the theoretical regret rate bound discussed in class, and to Agrawal and Goyal (2012)?

(d) Now let $k = 2$, fix $\theta^1 = .5$ and $T = 200$. Plot cumulative average regret for $T$ as a function of $\theta^2$, for $\theta^2 \in [0, 1]$. Do the same for the share of observations assigned to the optimal treatment.

How does the result relate to the local-to-zero asymptotics discussed in class, and to Figure 3 in Wager and Xu (2021)?

2. In this problem, we will again consider the Bernoulli bandit, and compare Thompson sampling to exploration sampling, as discussed in Kasy and Sautmann (2021).

(a) Create a modified version of the first function from problem 1, where instead of Thompson sampling treatment is assigned using exploration sampling.

Let this function additionally return the treatment $d^*_T$ with the highest posterior mean.

(b) Create a modified version of the second function from problem 1, again replacing Thompson sampling by exploration sampling. Exploration sampling assigns treatment $d$ with probability

$$q_t^d = \frac{p_t^d(1 - p_t^d)}{\sum_{d'} p_t^{d'}(1 - p_t^{d'})},$$

where $p_t^d$ is the posterior probability that treatment $d$ is optimal.s

Let this function additionally return the average across replications of policy regret

$$\left(\max_d \theta^d\right) - \theta^{d^*_T},$$

and the probability of choosing the best arm, $P(d^*_T = \text{argmax}_d\, \theta^d)$. Edit the second function from problem 1 to do the same for Thompson sampling.

(c) Pick a fixed vector of parameters $(\theta^1, \ldots, \theta^k)$ and a time horizon $T$ and calculate cumulative average regret as well as average policy regret, for both Thompson sampling and exploration sampling. Do so using a large number of replications $R$ (such as $R = 10.000$).

How does the result line up with the discussion and simulations of Kasy and Sautmann (2021)?

(d) Repeat this exercise for several different parameter vectors $(\theta^1, \ldots, \theta^k)$ and sample sizes $T$. Discuss any patterns you might find.

# References

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132.

Wager, S. and Xu, K. (2021). Diffusion asymptotics for sequential experiments. *arXiv preprint arXiv:2101.09855.*