Stochastic bandits with feedback graphs

Gurvan L'Hostis
gurvan.lhostis@polytechnique.edu

Ronan Riochet
ronan.riochet@polytechnique.edu

Abstract

We present and experiment with the algorithms of [1], which considers graphs with side observations and stochastic bandits, and those of [2] where directed graph are used in an adversarial context.

Our focus is to try and design algorithms for the general feedback graphs in a stochastic context; we present an adaptation of Thompson sampling as well as some algorithms that are inspired from UCB. Both our Thompson and UCB-derived methods produce good experimental results even though we did not derive theoretical upper bounds on the regret.

1 Introduction

1.1 Definitions

The problem we consider is derived from the classical *k-arms* bandit framework where a player chooses one of the k arms at each time step t and gets a reward $X_{i,t}$.

The performance of a strategy is measured by the difference between its rewards and that of an optimal strategy which is called the *regret*:

$$\mathbb{E}[R(n)] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)] \tag{1}$$

In our problem, we consider that the performance of the player is still only related to the rewards given by the arms she chooses, but we add the following hypotheses:

- The players observes rewards from the set of bandit according to a directed graph. For instance, when choosing arm i, the rewards of the out-neighbours of node i in the graph are observed.
- The reward of the chosen arm is not necessarily observed, but it is the one that matters for regret computations.

The name *general feedback graphs* is given to the class of problems in this setup, and the subclass of *side-observation graphs* is those where we add the condition that the chosen arm is always observed (all self-loops are present in the graph).

1.2 Graphs

There are three classes of graphs defined by [2]:

- Strongly observable: when every vertex either has a self-loop or all other vertices as in, regret is in $\Theta(T^{1/2})$;
- Unobservable: when there is at least one vertex with no in-neighbours, expected regret is in $\Theta(T)$;
- Weakly observable: when the graph is observable but not strongly, regret is in $\Theta(T^{2/3})$.

We will use the graphs that are suggested in [2], they are reported in figure 1. The first three are strongly observable and the last two are weakly observable.

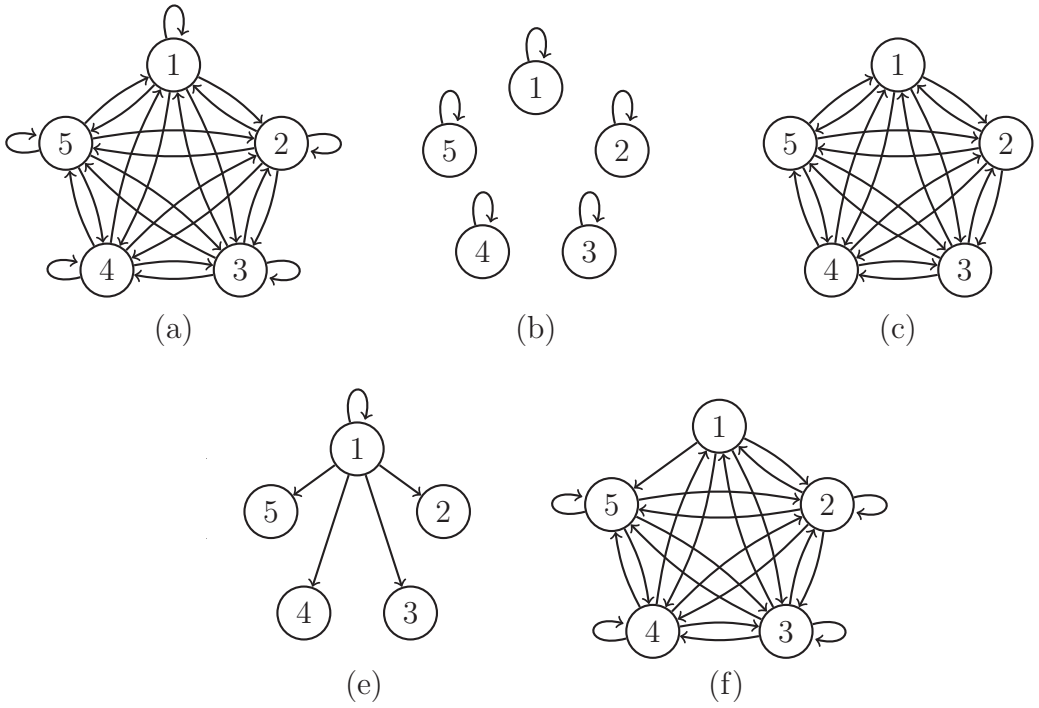

Figure 1: The five graphs we use for our experiments: (a) full, (b) bandit, (c) loopless, (d) revealing, (e) weak.

The bandits that we use are stochastic and have rewards in $[0, 1]$.

1.3 Algorithms

We will first focus on the UCB algorithms of [1] which are designed for stochastic problems with side-observation graphs and derive similar Thompson algorithms. Next, we will show how Exp3G from [2] proposes a solution for the general feedback graphs were the aforementioned strategies fail. We will then present a few strategies of our own that we tried with no theoretical guarantees.

In the following, we use the same notations as in [1]:

K	Number of arms
Δ_i	difference between the mean reward of optimal arm and arm i
$X_{i,t}$	reward of arm i
I_t	index of the arm player at t
N(i)	neighbourhood of arm i (includes i)
$O_i(n)$	observations for arm i after n steps

2 Stochastic Bandit algorithms for side-observation graphs

In this section we present some algorithms for the special case of *side-observation graphs* were the decision maker always observes her reward. The paper [1] proposes two algorithms for this problem, based on upper confidence bounds. In this section, we will first present those two algorithms with their associated results. We will then present two algorithms that are adapted from Thompson sampling with empirical results.

2.1 UCB algorithms

Two UCB-based algorithms are presented in [1]: UCB-N (see algorithm 1) and UCB-MaxN (see algorithm 2). We call UCB for arm i at time t, the quantity:

$$\text{UCB}_i(t) = \bar{X}_i + \sqrt{\frac{2\log t}{O_t}}$$

Where $\sqrt{\frac{2\log t}{O_t}}$ can be interpreted as a *bias term* to the sample mean of arm i. It is proven by [3] that picking $\text{argmax}_i \text{UCB}_(t)$ at each time achieves to the following upper bound:

$$\mathbb{E}[R(n)] \leq 8\big(\sum_{i=1}^{K} \frac{1}{\Delta_i}\big)\log n + \big(1+\frac{\pi^2}{3}\big)\sum_{i=1}^{K}\Delta_i \tag{2}$$

We expect that in the case of a side-observation graph, the decision-maker can benefit from this additional information to decrease her regret. In [1], authors prove that the UCB-N algorithm improves this upper bound on side-observation graphs, as in the following theorem.

Theorem 1 *The expected regret of policy UCB-N after n steps is upper bounded by:*

$$\mathbb{E}[R(n)] \leq \inf_{\mathcal{C}} \big\{8\big(\sum_{C\in\mathcal{C}} \frac{\max_{i\in C}\Delta_i}{\Delta_C^2}\big)\log n\big\} + \big(1+\frac{\pi^2}{3}\big)\sum_{i=1}^{K}\Delta_i \tag{3}$$

where $\Delta_C = \min_{i\in\mathcal{C}}\Delta_i$.

Hence, UCB-N reduces the logarithmic factor in (2). In equation (3) however, the regret is in $\mathcal{O}(K)$. One can have the intuition that side information should lower this quantity. Indeed, in a side-observation graph, pulling an arm i gives the same information as pulling every arm of its out-neighbourhood in a bandit problem. The following theorem confirms this idea, proving that this term can be reduced to $\mathcal{O}(|\mathcal{C}|)$.

Theorem 2 *The expected regret of policy UCB-MaxN after n steps is upper bounded by:*

$$\mathbb{E}[R(n)] \leq \inf_{\mathcal{C}} \big\{8\big(\sum_{C\in\mathcal{C}} \frac{\max_{i\in C}\Delta_i}{\Delta_C^2}\big)\log n\big\} + \big(1+\frac{\pi^2}{3}\big)\sum_{C\in\mathcal{C}}\Delta_i + o_{n\to 1}(1) \tag{4}$$

Data: Graph G

$\bar{X}, O \leftarrow 0, 0$
for $t \geq 1$ **do**
  $I \leftarrow \text{argmax}_i \{\bar{X}_i + \sqrt{\frac{2 \log t}{O_i}}\}$
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $O_k \leftarrow O_k + 1$
    $\bar{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\bar{X}_k$
  end
end

Algorithm 1: UCB-N

Data: Graph G

$\bar{X}, O \leftarrow 0, 0$
for $t \geq 1$ **do**
  $J \leftarrow \text{argmax}_j \{\bar{X}_i + \sqrt{\frac{2 \log t}{O_i}}\}$
  $I \leftarrow \text{argmax}_{i \in N^{\text{in}}(J)} (\bar{X}_{i,t})$
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $O_k \leftarrow O_k + 1$
    $\bar{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\bar{X}_k$
  end
end

Algorithm 2: UCB-MaxN

2.2 Thompson algorithms

Our first attempts at developing our own algorithms consist in adapting Thomson Sampling [4] to side-observation graphs just like UCB is adapted in [1].

We also sample from Beta Bayesian priors on the arm means but

- we use all the feedback provided by the graph to update our priors in both *Thompson-N* and *Thompson-MaxN*;

- in *Thompson-MaxN*, we draw the arm which gives an observation of the sampled arm with highest mean reward, which is equal to $\frac{S}{S+F}$.

Data: Graph G

$S, F \leftarrow 0, 0$
for $t \geq 1$ **do**
  $\theta \leftarrow \text{Beta}(S+1, F+1)$
  $I \leftarrow \text{argmax}_i(\theta)$
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $S_k \leftarrow S_k + X_{k,t}$
    $F_k \leftarrow F_k + (1 - X_{k,t})$
  end
end

Algorithm 3: Thompson-N

Data: Graph G

$S, F \leftarrow 0, 0$
for $t \geq 1$ **do**
  $\theta \leftarrow \text{Beta}(S+1, F+1)$
  $J \leftarrow \text{argmax}_j(\theta)$
  $I \leftarrow \text{argmax}_{i \in N^{\text{in}}(J)}(\frac{S}{S+F})$
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $S_k \leftarrow S_k + X_{k,t}$
    $F_k \leftarrow F_k + (1 - X_{k,t})$
  end
end

Algorithm 4: Thompson-MaxN

3 Adversarial algorithm for general feedback graph

3.1 Exp3G

In introduction, we presented two results on weakly and strongly observable graphs:

- For strongly observable [1]: regret is in $\Theta(T^{1/2})$;
- For weakly observable [2]: regret is in $\Theta(T^{2/3})$.

In [2], authors present an algorithm, called EXP3.G (see figure 2) which achieves both of these upper bounds. EXP3.G proposes a trade-off between exploration and exploitation. The parameter γ defines the probability of doing an exploration move, that is to say pulling an arm uniformly in a given exploration set U. When doing a exploitation move, decision is made according to a probability distribution q_t, described below.

Exp3.G uses importance sampling to construct unbiased loss estimates $\hat{\ell}_t(i), i = 1...K$. First, it computes the probability $\mathbb{P}_t(i) = \mathbb{P}(i \in N^{out}(I_t))$ of observing the loss $\ell_i(t)$ upon playing $I_t \sim p_t$:

$$\mathbb{P}_t(i) = \sum_{j \in N^{in}(i)} p_t(j) \tag{5}$$

[1] when every vertex either has a self-loop or all other vertices as in-neighbours
[2] when the graph is observable but not strongly

followed by the loss estimates:

$$\hat{\ell}_t(i) = \frac{\ell_t(i)}{\mathbb{P}_t(i)} \mathbb{I}\{i \in N^{out}(I_t)\} \tag{6}$$

(Note that this estimator is unbiased and with controlled variance $\mathbb{E}_t[\hat{\ell}_t(i)^2] = \frac{\ell_t(i)^2}{\mathbb{P}_t(i)}$). Finally the q_{t+1} is updated as follow:

$$q_{t+1}(i) \propto q_t(i) \exp(-\eta \hat{\ell}_t(i)),\ \forall i \in V \tag{7}$$

A pseudo-code of the Exp3.G is given in Figure 2.

Parameters: Feedback graph $G = (V, E)$, learning rate $\eta > 0$,
exploration set $U \subseteq V$, exploration rate $\gamma \in [0, 1]$

Let u be the uniform distribution over U;
Initialize q_1 to the uniform distribution over V;
For round $t = 1, 2, \ldots$
Compute $p_t = (1 - \gamma) q_t + \gamma u$;
Draw $I_t \sim p_t$, play I_t and incur loss $\ell_t(I_t)$;
Observe $\{(i, \ell_t(i)) : i \in N^{\text{out}}(I_t)\}$;
Update

$$\forall\, i \in V \qquad \widehat{\ell}_t(i) \;=\; \frac{\ell_t(i)}{P_t(i)}\, \mathbb{I}\{i \in N^{\text{out}}(I_t)\}\,, \qquad \text{with} \qquad P_t(i) \;=\; \sum_{j \in N^{\text{in}}(i)} p_t(j)\;; \tag{1}$$

$$\forall\, i \in V \qquad q_{t+1}(i) \;=\; \frac{q_t(i) \exp(-\eta \widehat{\ell}_t(i))}{\sum_{j \in V} q_t(j) \exp(-\eta \widehat{\ell}_t(j))}\;; \tag{2}$$

Figure 2: Exp3.G - online learning with a feedback graph.

The following theorem ([2]) proves that Exp3.G algorithm achieves upper bounds presented above.

Theorem 3 *Let G=(V,E) be a feedback graph with $K = |V|$, independence number $\alpha = G(\alpha)$ and weakly dominating number $\delta = \delta(G)$. Let D be a weakly dominating set such that $|D| = \delta$. The expected regret of Algorithm (2) on the online learning problem induced by G satisfies the following:*

- *if G is strongly observable, then for U=V, $\gamma = \min\left\{\left(\frac{1}{\alpha T}\right)^{1/2}, \frac{1}{2}\right\}$ and $\eta = 2\gamma$, the expected regret against any loss sequence is $\mathcal{O}(\alpha^{\frac{1}{2}} T^{\frac{1}{2}} \log KT)$;*
- *if G is weakly observable and $T \geq K^3 ln(K)/\delta^2$, then for U=V, $\gamma = \min\left\{\left(\frac{\delta \log K}{T}\right)^{\frac{1}{3}}, \frac{1}{2}\right\}$ and $\eta = \frac{\gamma^2}{\delta}$, the expected regret against any loss sequence is $\mathcal{O}((\delta \log K)^{\frac{1}{3}} T^{\frac{2}{3}})$.*

4 Experiments

Experiments with the five aforementioned strategies are presented in figure 3. We show the average regret curves for two bandit mean configurations:

- Decreasing: $[5/6, 4/6, 3/6, 2/6, 1/6]$;
- Increasing: $[1/6, 2/6, 3/6, 4/6, 5/6]$;

Graphs *full*, *bandit* and *loopless* are symmetric and the order of means should not matter. However, it is important to show two configurations where arm 1 either is the optimum arm or not because it has a specific role in graphs *revealing* and *weak*.

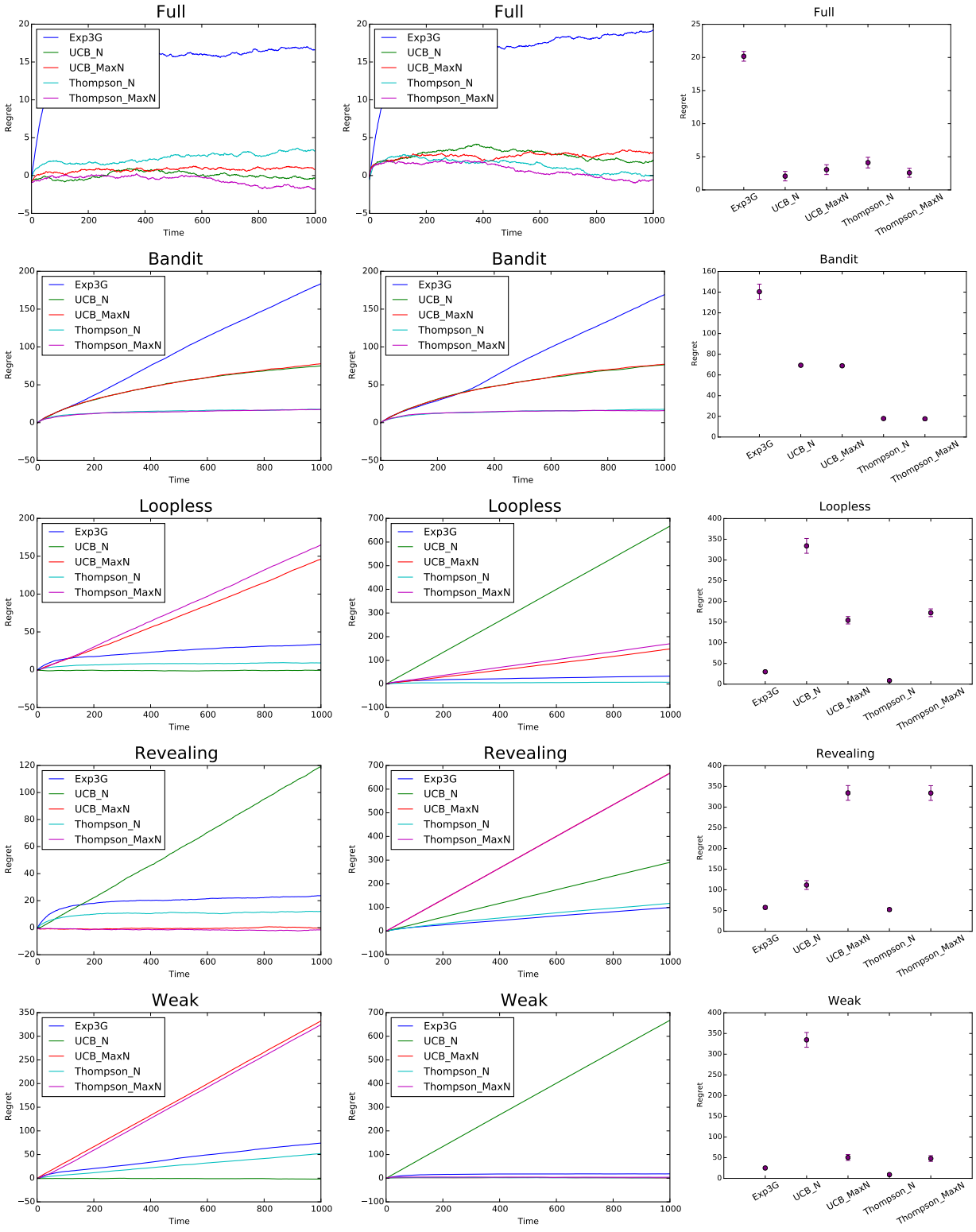

Figure 3: Experiments with adversarial strategy Exp3G and stochastic strategies UCB-N, UCB-MaxN, Thompson-N and Thompson-MaxN on the five graphs. The left column is the average regret curve over 100 runs with arms of *decreasing* mean (1 has highest mean); the middle column is the average regret curve over 100 runs with arms of *increasing* mean (1 has lowest mean); the right column gives regret at $T = 1000$ averaged over 1000 bandit mean configurations drawn uniformly at random with 95% confidence intervals.

4.1 Side-observation graphs

Full All algorithms perform equally on the full graph with a regret that seems to be in $\mathcal{O}(1)$, Exp3G having a slight penalty in the beginning.

Bandit On the bandit graph however, we remark two intersting points. As can be expected, N/MaxN couples of strategies perform equally as there is no side observation to give an advantage to MaxN. All four stochastic methods produce logarithmic regret, but Thompson strategies achieve a better regret in all configurations. Exp3G has a regret that is clearly not logarithmic, which could be expected ([2] predicts a $\Theta(T^{1/2})$ regret for such a graph). It does not take advantage of the stochastic properties of the bandits.

4.2 General feedback graphs

Loopless On the loopless graph, it is interesting to remark that the performance of UCB-N changes between the two configurations even though they were supposed to be symmetric problems. The absence of self-loops makes the strategy function oddly, it gets stuck in the first arm it observes and there is no first round of observing everything (arm 1 is never observed and its bias term stays infinite).

Quite surprisingly, Thompson-N gives excellent performance, even better than Exp3G in general. We interpret this with the fact that Thompson algorithms are not optimistic and thus the lack of information when we exploit an arm does not cause the same problems as in UCB, where the optimistic sampling is not corrected over time in the absence of self-loop.

Revealing Here, the difference between the increasing and the decreasing configurations is widespread. The MaxN strategies will always draw 1 and are thus out of the question. UCB-N will always draw 2 whose bias term is infinite just like that of 1 in the loopless graph.

Thompson-N also performs better than Exp3G on average in this configuration though the difference is very small.

Weak This weakly observable graph does not give a confirmation of the $\Theta(T^{2/3})$ theoretical regret.

The same troubles as in the other general feedback graph appear for stochastic methods, however they penalize the MaxN methods only when 1 has highest mean so they dont do too poorly on average.

Thompson-N still beats Exp3G in this setup.

5 Attempts at generalising stochastic algorithms

Exp3G is a valid algorithm even in the case where there is no self observation but it is under performing in the context of stochastic bandits because it does not use that assumption. We therefore try to design algorithms that are better-suited for the case of general feedback graphs with stochastic bandits.

5.1 Generalising UCB

Our approach is to start from the UCB algorithm and modify it so that it works in the general feedback case. The problem with UCB-N is that it can get stuck with an under-performing arm because the estimate of its mean is wrong and is not updated, and the problem with UCB-MaxN is that it will tend to choose the best in-neighbour of the arm that has the highest mean, which is consistently under-performing if we there is no self-observation of the best arm.

In a word, the exploitation-exploration trade-off cannot be dealt with just the "information need" bias anymore. Our approach therefore consists in trying to restore that trade-off.

Augmenting the space of actions The first attempt consists in doubling the the space of possible actions at every iteration. Instead of using $\bar{X} + (2 * \log(T)/O)^{1/2}$, we put each arm in the set of choices twice with weights $\bar{X}$ and $\alpha\,(2 * \log(T)/O)^{1/2}$.

With this space of choices, the strategy chooses either an exploitation move when the highest weight is a mean estimate or an exploration move when the highest is a "bias". The trade-off is made explicit; we add a hyperparameter α to regulate it.

Data: Graph G, trade-off α

$\bar{X}, O \leftarrow 0, 0$
for $t \geq 1$ **do**
  if $\max_i \bar{X} \geq \alpha \max_i \sqrt{\frac{2 \log t}{O_i}}$ **then**
    $I \leftarrow \text{argmax}_i \bar{X}_i$
  else
    Find the less observed vertice (lov):
    lov $\leftarrow \text{argmin}_i O_i$
    $I \leftarrow \text{argmax}_{i \in N^{\text{in}}(\text{lov})} \bar{X}_i$
  end
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $O_k \leftarrow O_k + 1$
    $\bar{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\bar{X}_k$
  end
end

Algorithm 5: Separate-UCB

Q-Learning style exploration The approach of having a double-space of exploration/exploitation moves boils down to regularly exploring the unobserved arms in practice. It is thus interesting to try and make that regular exploration explicit with an exploration strategy akin to what is done in *Q-learning*.

We tested two such strategies with respectively an exploration rate decreasing as an inverse function of the number of iterations and another with exponential decrease. Both are parametrised by α.

Data: Graph G, exploration parameter $\alpha > 0$

$\bar{X}, O \leftarrow 0, 0$
for $t \geq 0$ **do**
  Choose between exploitation and exploration:
  $do\_explo = \text{Bernoulli}(\frac{1}{1+\alpha t})$
  if *do_explo* **then**
    $J \leftarrow \text{argmax}_i \{\bar{X}_i + \sqrt{\frac{2 \log t}{O_i}}\}$
    $I \leftarrow \text{argmax}_{i \in N^{\text{in}}(\text{J})} \bar{X}_i$
  else
    $I \leftarrow \text{argmax}_i \bar{X}_{i,t}$
  end
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $O_k \leftarrow O_k + 1$
    $\bar{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\bar{X}_{k,t}$
  end
end

Algorithm 6: Q-UCB-inv

Data: Graph G, exploration parameter $\alpha > 0$

$\bar{X}, O \leftarrow 0, 0$
for $t \geq 0$ **do**
  Choose between exploitation and exploration:
  $do\_explo = \text{Bernoulli}(\exp(-\alpha t))$
  if *do_explo* **then**
    $J \leftarrow \text{argmax}_i \{\bar{X}_i + \sqrt{\frac{2 \log t}{O_i}}\}$
    $I \leftarrow \text{argmax}_{i \in N^{\text{in}}(\text{J})} \bar{X}_i$
  else
    $I \leftarrow \text{argmax}_i \bar{X}_{i,t}$
  end
  PULL arm I and observe feedbacks X_t
  for $k \in N^{out}(I)$ **do**
    $O_k \leftarrow O_k + 1$
    $\bar{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\bar{X}_{k,t}$
  end
end

Algorithm 7: Q-UCB-exp

5.2 Experiments

We compare the strategies of this section to Exp3G and UCB-N in a similar fashion as with previous experiments, see figure 4.

Just like before, the three new algorithms perform well on the full graph.

Q-UCB-inv We observe that Q-UCB-inv performs similarly to the MaxN algorithms, and by observing its regret curves on *revealing* and *weak* we can deduce that it has the same problems caused by drawing the max of the in-neighbours. In a word, it does not learn fast enough and is penalised by too much exploration.

Separate-UCB and Q-UCB-exp achieve similar performance, they are slightly better than Thompson-N on general feedback graphs and Thompson-N is better in the *bandit* setting. Q-UCB-exp would be the strategy to retain as it achieves best performance of the three and is more explicit than Separate-UCB.

6 Conclusion

With UCB as a starting point, we derived an algorithm that uses the stochastic assumptions and is adapted to general feedback graphs by making the exploration-exploitation trade-off explicit. This algorithm beats both UCB and Exp3G strategies in all studied graphs.

However, the strategy that produced the best results on the whole was one that we implemented with only the side-observation setup in mind: Thompson-N. While UCB fails without self-loops because it needs them to correct its optimistic biases, Thompson sampling does not get stuck and explores all arms appropriately. It would now be interesting to derive regret upper bounds for it in the general feedback graphs context.

References

[1] Stephane Caron et al. "Leveraging Side Observations in Stochastic Bandits." In: *UAI*. Ed. by Nando de Freitas and Kevin P. Murphy. AUAI Press, 2012, pp. 142–151. URL: http://dblp.uni-trier.de/db/conf/uai/uai2012.html#CaronKLB12.

[2] Noga Alon et al. "Online Learning with Feedback Graphs: Beyond Bandits". In: *CoRR* (2015). URL: http://arxiv.org/abs/1502.07617.

[3] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Mach. Learn.* 47.2-3 (May 2002), pp. 235–256.

[4] Shipra Agrawal and Navin Goyal. "Analysis of Thompson Sampling for the Multi-armed Bandit Problem." In: *COLT*. Ed. by Shie Mannor, Nathan Srebro, and Robert C. Williamson. Vol. 23. JMLR Proceedings. 2012, pp. 39.1–39.26.

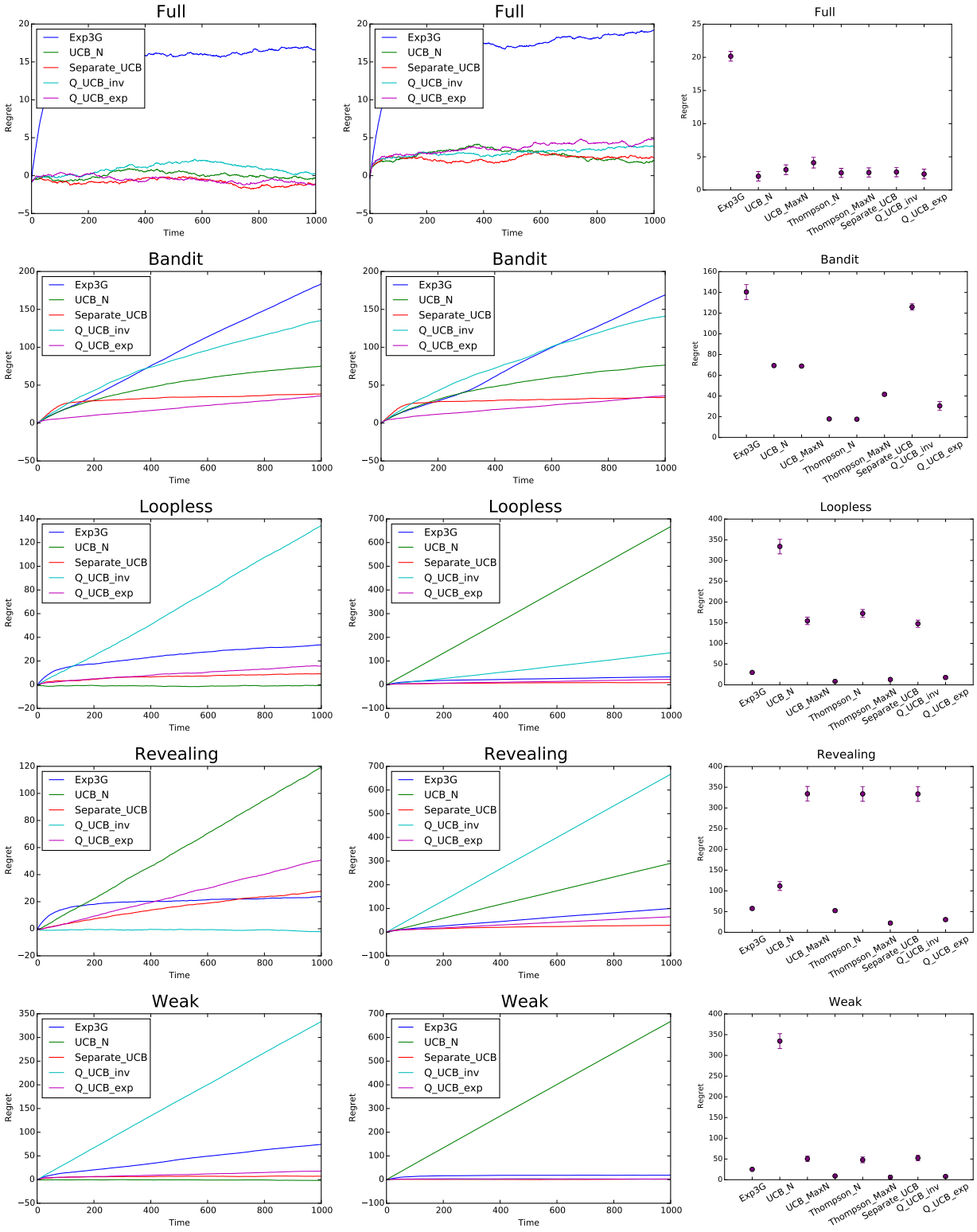

Figure 4: Experiments with adversarial/general feedback strategy Exp3G, stochastic/side-observation strategy UCB-N and stochastic/general feedback strategies Separate-UCB, Q-UCB-inv and Q-UCB-exp on the five graphs. The left column is the average regret curve over 100 runs with arms of *decreasing* mean (1 has highest mean); the middle column is the average regret curve over 100 runs with arms of *increasing* mean (1 has lowest mean); the right column gives regret at $T = 1000$ averaged over 1000 bandit mean configurations drawn uniformly at random with 95% confidence intervals for all eight strategies presented in the report.