This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at http://banditalgs.com. Feedback on any aspect is very welcome: banditalgs@gmail.com

31 Non-Stationary Bandits

The competitor class used in the standard definition of the regret is not appropriate when the underlying environment is changing. In this chapter we increase the power of the competitor class to 'track' changing environments and derive algorithms for which the regret relative to this enlarged class is not too large. While we specify the results for bandits with finitely many arms (both stochastic and adversarial), many of the ideas generalize to other models such as linear bandits. This chapter also illustrates the flexibility of the tools presented in the earlier chapters, which are applied here almost without modification. We hope (and expect) that this will also be true for other models you might study.

31.1 Adversarial bandits

In contrast to stochastic bandits, the adversarial bandit model presented in Chapter 11 does not prevent the environment from changing over time. The problem is that bounds on the regret can become vacuous when the losses appear nonstationary. To illustrate an extreme situation, suppose you face a two-armed adversarial bandit with losses $y_{t1} = \mathbb{I}\{t \leq n/2\}$ and $y_{t2} = \mathbb{I}\{t > n/2\}$. If we run Exp3 on this problem, then Theorem 11.2 guarantees that

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} y_{tA_t}\right] - \min_{i \in \{1,2\}} \sum_{t=1}^{n} y_{ti} \leq \sqrt{2nK \log(K)}\,.$$

Since $\min_{i \in \{1,2\}} \sum_{t=1}^{n} y_{ti} = n/2$, by rearranging we see that

$$\mathbb{E}\left[\sum_{t=1}^{n} y_{tA_t}\right] \leq \frac{n}{2} + \sqrt{2nK \log(K)}\,.$$

To put this in perspective, a policy that plays each arm with probability half in every round would have $\mathbb{E}[\sum_{t=1}^{n} y_{tA_t}] = n/2$, for *any* sequence $(y_{ta})_{ta}$ of losses in the $[0, 1]$ interval. In other words, the regret guarantee is practically meaningless.

What should we expect for this problem? The sequence of losses is so regular that we might hope that a clever policy will mostly play the second arm in the first $n/2$ rounds and then switch to playing mostly the first arm in the second $n/2$ rounds. Then the cumulative loss would be close to zero and the regret would be negative. Rather than aiming to guarantee negative regret, we can redefine the

regret by enlarging the competitor class as a way to ensure meaningful results. Let $\Gamma_{n,m} \subset [K]^n$ be the set of action sequences of length n with at most $m-1$ changes:

$$\Gamma_{n,m} = \left\{ (a_t) \in [K]^n : \sum_{t=1}^{n-1} \mathbb{I}\{a_t \neq a_{t+1}\} \leq m-1 \right\}.$$

Then define the **nonstationary regret** with $m-1$ change points by

$$R_{n,m} = \mathbb{E}\left[\sum_{t=1}^{n} y_{tA_t}\right] - \min_{a \in \Gamma_{n,m}} \mathbb{E}\left[\sum_{t=1}^{n} y_{ta_t}\right].$$

The nonstationary regret is sometimes called the **tracking regret** because a learner that makes it small must 'track' the best arm as it changes. Notice that $R_{n,1}$ coincides with the usual definition of the regret. Furthermore, on the sequence described at the beginning of the section we see that

$$R_{n,2} = \mathbb{E}\left[\sum_{t=1}^{n} y_{tA_t}\right],$$

which means a policy can only enjoy sublinear nonstationary regret if it detects the change point quickly. The obvious question is whether or not such a policy exists and how its regret depends on m.

Exp4 for nonstationary bandits

One idea is to use the Exp4 policy from Chapter 18 with a large set of experts, one for each $a \in \Gamma_{n,m}$. Theorem 18.1 shows that Exp4 with these experts suffers regret of at most

$$R_{n,m} \leq \sqrt{2nK \log |\Gamma_{n,m}|}. \tag{31.1}$$

Naively bounding $\log |\Gamma_{n,m}|$ (Exercise 31.1) and ignoring constant factors shows that

$$R_{n,m} = O\left(\sqrt{nmK \log\left(\frac{Kn}{m}\right)}\right). \tag{31.2}$$

To see that you cannot do much better than this, imagine interacting with m adversarial bandit environments sequentially, each with horizon n/m. No matter what policy you propose, there exist choices of bandits such that the expected regret suffered against each bandit is at least $\Omega(\sqrt{nK/m})$. And after summing over the m instances we see that the worst case regret is at least

$$R_{n,m} = \Omega\left(\sqrt{nmK}\right),$$

which matches the upper bound except for logarithmic factors. Notice how this lower bound applies to policies that know the location of the changes, so it is not true that things are significantly harder in the absence of this knowledge. There is one big caveat with all these calculations. The running time of a naive

implementation of Exp4 is linear in the number of experts, which even for modestly sized m is very large indeed.

Online Stochastic Mirror Descent

The computational issues faced by Exp4 are most easily overcome using the tools from online convex optimization developed in Chapter 28. The idea is to use online stochastic mirror descent and the negentropy potential. Without further modification this would be Exp3, which you will show does not work for nonstationary bandits (Exercise 31.3). The trick is to restrict the action set to the clipped simplex $\mathcal{A} = \mathcal{P}_{K-1} \cap [\alpha, 1]^K$ where $\alpha \in [0, 1/K]$ is a constant to be tuned subsequently. The clipping ensures the algorithm does not commit too hard to any single arm. The rationale is that a strong commitment could prevent this discovery of change points.

Let $F : [0, \infty)^K \to \mathbb{R}$ be the unnormalized negentropy potential and $P_1 \in \mathcal{A}$ be the uniform probability vector. In each round t the learner samples $A_t \sim P_t$ and updates its sampling distribution using

$$P_{t+1} = \operatorname{argmin}_{p \in \mathcal{A}} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t)\,, \tag{31.3}$$

where $\eta > 0$ is the learning rate and $\hat{Y}_{ti} = \mathbb{I}\{A_t = i\}\, y_{ti}/P_{ti}$ is the importance-weighted estimator of the loss of action i for round t. The solution to the optimization problem of Eq. (31.3) can be computed efficiently using the two-step process:

$$\begin{aligned} \tilde{P}_{t+1} &= \operatorname{argmin}_{p \in [0,\infty)^K} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t)\,, \\ P_{t+1} &= \operatorname{argmin}_{p \in \mathcal{A}} D_F(p, \tilde{P}_{t+1})\,. \end{aligned}$$

The first of these subproblems can be evaluated analytically, yielding $\tilde{P}_{t+1,i} = P_{ti} \exp(-\eta \hat{Y}_{ti})$. The second can be solved efficiently using the result in Exercise 26.9. The algorithm enjoys the following guarantee on its regret:

THEOREM 31.1 *The expected regret of the policy sampling $A_t \sim P_t$ with P_t defined in Eq. (31.3) is bounded by*

$$R_{n,m} \le \alpha n(K-1) + \frac{m \log(1/\alpha)}{\eta} + \frac{\eta n K}{2}\,.$$

Proof Let $a^* \in \operatorname{argmin}_{a \in \Gamma_{n,m}} \sum_{t=1}^n y_{ta_t}$ be an optimal sequence of actions in hindsight constrained to $\Gamma_{n,m}$. Then let $1 = t_1 < t_2 < \cdots < t_m < t_{m+1} = n$ so that a^*_t is constant on each interval $\{t_i, \ldots, t_{i+1} - 1\}$. We abuse notation by writing $a^*_i = a^*_{t_i}$. Then the regret decomposes into

$$\begin{aligned} R_{n,m} &= \mathbb{E}\left[\sum_{t=1}^n (y_{tA_t} - y_{ta^*_t})\right] = \mathbb{E}\left[\sum_{i=1}^m \sum_{t=t_i}^{t_{i+1}-1} (y_{tA_t} - y_{ta^*_t})\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}-1} (y_{tA_t} - y_{ta^*_i}) \,\middle|\, P_{t_i}\right]\right]. \end{aligned}$$

The next step is to apply Eq. (28.10) and the solution to Exercise 28.5 to bound the inner expectation, giving

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}-1}(y_{tA_t}-y_{ta_i^*})\,\middle|\,P_{t_i}\right] &= \mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\langle P_t - e_{a_i^*}, y_t\rangle\,\middle|\,P_{t_i}\right] \\ &\le \alpha(t_{i+1}-t_i)(K-1) + \mathbb{E}\left[\max_{p\in\mathcal{A}}\sum_{t=t_i}^{t_{i+1}-1}\langle P_t - p, y_t\rangle\,\middle|\,P_{t_i}\right] \\ &\le \alpha(t_{i+1}-t_i)(K-1) + \mathbb{E}\left[\max_{p\in\mathcal{A}}\frac{D(p,P_{t_i})}{\eta} + \frac{\eta K(t_{i+1}-t_i)}{2}\,\middle|\,P_{t_i}\right].\end{aligned}$$

By assumption, $P_{t_i} \in \mathcal{A}$ and so $P_{t_ij} \ge \alpha$ for all j and $D(p, P_{t_i}) \le \log(1/\alpha)$. Combining this observation with the previous two displays shows that

$$R_{n,m} \le n\alpha(K-1) + \frac{m\log(1/\alpha)}{\eta} + \frac{\eta nK}{2}\,. \qquad \square$$

The learning rate and clipping parameters are approximately optimized by

$$\eta = \sqrt{2m\log(1/\alpha)/(nK)} \qquad \text{and} \qquad \alpha = \sqrt{m/(nK)}\,,$$

which leads to a regret of $R_{n,m} \le \sqrt{mnK\log(nK/m)} + \sqrt{mnK}$. In typical applications the value of m is not known. In this case one can choose $\eta = \sqrt{\log(1/\alpha)/nK}$ and $\alpha = \sqrt{1/nK}$ and the regret increases by a factor of $O(\sqrt{m})$.

31.2 Stochastic bandits

We saw in Part II that by making a statistical assumption on the rewards it was possible to design policies with logarithmic instance-dependent regret. This is the big advantage of making assumptions – you get stronger results. The nonstationarity makes the modelling problem less trivial. To keep things simple we will assume the rewards are Gaussian and that for each arm i there is a function $\mu_i : [n] \to \mathbb{R}$ and the reward is

$$X_t = \mu_{A_t}(t) + \eta_t\,,$$

where (η_t) is a sequence of independent standard Gaussian random variables. The optimal arm in round t has mean $\mu^*(t) = \max_{i\in[K]}\mu_i(t)$ and the regret is

$$R_n(\mu) = \sum_{t=1}^n \mu^*(t) - \mathbb{E}\left[\sum_{t=1}^n \mu_{A_t}(t)\right].$$

The amount of nonstationarity is modelled by placing restrictions on the functions $\mu_i : [n] \to \mathbb{R}$. To be consistent with the previous section we assume the mean vector changes at most $m-1$ times, which amounts to saying that

$$\sum_{t=1}^{n-1}\max_{i\in[K]}\mathbb{I}\left\{\mu_i(t) \ne \mu_i(t+1)\right\} \le m-1\,.$$

If the locations of the change points were known then running a new copy of UCB on each interval would lead to a bound of

$$R_n(\mu) = O\left(\frac{mK}{\Delta_{\min}} \log\left(\frac{n}{m}\right)\right),$$

where $\Delta_{\min}$ is the smallest suboptimality gap over all m blocks. In the last section we saw that the bound achieved by an omniscient policy that knows when the changes occur can be achieved by a policy that does not. Unfortunately this is not true here.

THEOREM 31.2 *Let $K = 2$ and suppose that $\mu_i(t) = \mu_i$ is constant for both arms and $\Delta = \mu_1 - \mu_2 > 0$. Let π be a policy such that its expected regret $R_n(\mu)$ on bandit μ satisfies $R_n(\mu) = o(n^\alpha)$ with $0 < \alpha \leq 1$. Then, for all sufficiently large n there exists a nonstationary bandit μ' with at most two change points and $\min_{t\in[n]} |\mu'_1(t) - \mu'_2(t)| \geq \Delta$ such that $R_n(\mu') \geq cn/R_n(\mu)$, where $c > 0$ is a constant that depends only on $\alpha > 0$.*

The theorem shows that if a policy enjoys $R_n(\mu) = o(n^{1/2})$ for any nontrivial (stationary) bandit, then its minimax regret is at least $\omega(n^{1/2})$ on some nonstationary bandit. In particular, if $R_n(\mu) = O(\log(n))$, then the minimax regret is at least $\Omega(n/\log(n))$. This dashes our hopes for a policy that is much better than Exp4 in a stochastic setting. There are algorithms designed for nonstationary bandits in the stochastic setting with abrupt change points as described above. Those that come with theoretical guarantees are based on forgetting or discounting data so that decisions of the algorithm depend almost entirely on recent data. In the notes we discuss these approaches along with alternative models for nonstationarity.

Proof of Theorem 31.2 Let $(S_k)_{k=1}^L$ be a partition of $[n]$ to be specified later. Let $\mathbb{P}$ and $\mathbb{E}[\cdot]$ denote the probabilities and expectations with respect to the bandit determined by μ and $\mathbb{P}'$ with respect to alternative nonstationary bandit μ' to be defined shortly. By the pigeonhole principle there exists a $k \in [L]$ such that

$$\mathbb{E}\left[\sum_{t\in S_k} \mathbb{I}\{A_t = 2\}\right] \leq \frac{\mathbb{E}[T_2(n)]}{L}.$$

Define an alternative nonstationary bandit with $\mu'(t) = \mu$ except for $t \in S_k$ when we let $\mu'_2(t) = \mu_2 + \varepsilon$ where $\varepsilon = \sqrt{2L/\mathbb{E}[T_2(n)]}$. Then by Lemma 15.1 and Theorem 14.2,

$$\begin{aligned}
\mathbb{P}\left(\sum_{t\in S_k} \mathbb{I}\{A_t = 2\} \geq \frac{|S_k|}{2}\right) + \mathbb{P}'\left(\sum_{t\in S_k} \mathbb{I}\{A_t = 2\} < \frac{|S_k|}{2}\right) &\geq \frac{1}{2}\exp\left(-\mathrm{D}(\mathbb{P}, \mathbb{P}')\right) \\
&\geq \frac{1}{2}\exp\left(-\frac{\mathbb{E}[T_2(n)]\varepsilon^2}{2L}\right) \geq \frac{1}{2e}.
\end{aligned}$$

By Markov's inequality,

$$\mathbb{P}\left(\sum_{t\in S_k}\mathbb{I}\{A_t=2\}\geq\frac{|S_k|}{2}\right)\leq\frac{2}{|S_k|}\mathbb{E}\left[\sum_{t\in S_k}\mathbb{I}\{A_t=2\}\right]\leq\frac{2\mathbb{E}[T_2(n)]}{L|S_k|}\leq\frac{1}{\Delta^2|S_k|}\,,$$

where the last inequality follows by choosing $L=\left\lceil 2\Delta^2\mathbb{E}[T_2(n)]\right\rceil$, which also ensures that $\varepsilon-\Delta\geq\varepsilon/2$. Therefore,

$$R_n(\mu')\geq\left(\frac{1}{2e}-\frac{1}{\Delta^2|S_k|}\right)\frac{\varepsilon|S_k|}{4}\geq\left(\frac{1}{2e}-\frac{1}{\Delta^2|S_k|}\right)\frac{|S_k|\Delta}{2}\,.$$

If (S_k) is chosen as a uniform partition so that $|S_k|\geq\lfloor n/L\rfloor$, then, using $R_n(\mu)\geq\Delta\mathbb{E}\left[T_2(n)\right]$, the definition of L and that by assumption $R_n(\mu)=o(n^\alpha)$, we get that there exists a universal constant $c>0$ that depends on α only such that for sufficiently large n, $R_n(\mu')\geq cn/R_n(\mu)$. □

31.3 Notes

1 Environments that appear nonstationary can often be made stationary by adding context. For example, when bandit algorithms are used for on-line advertising, gym membership advertisements are received more positively in January than July. A bandit algorithm that is oblivious to the time of year will perceive this environment as nonstationary. You could tackle this problem by using one of the algorithms in this chapter. Or you could use a contextual bandit algorithm and include the time of year in the context. The reader is encouraged to consider whether or not adding contextual information might be preferable to using an algorithm designed for nonstationary bandits.

2 The negative results for stochastic nonstationary bandits do not mean that trying to improve on the adversarial bandit algorithms is completely hopeless. First of all, the adversarial bandit algorithms are not well suited for exploiting distributional assumptions on the noise, which makes things irritating when the losses/rewards are Gaussian (which are unbounded) or Bernoulli (which have small variance near the boundaries). There have been several algorithms designed specifically for stochastic nonstationary bandits. When the reward distributions are permitted to change abruptly as in the last section, then the two main algorithms are based on the idea of 'forgetting' rewards observed in the distant past. One way to do this is with **discounting**. Let $\gamma\in(0,1)$ be the **discount factor** and define

$$\hat{\mu}_i^\gamma(t)=\sum_{s=1}^{t}\gamma^{t-s}\mathbb{I}\{A_s=i\}\,X_s\qquad\qquad T_i^\gamma(t)=\sum_{s=1}^{t}\gamma^{t-s}\mathbb{I}\{A_s=i\}\,.$$

Then, for appropriately tuned constant α, the Discounted UCB policy chooses

each arm once and subsequently

$$A_t = \operatorname{argmax}_{i \in [K]} \left(\hat{\mu}_i^\gamma(t-1) + \sqrt{\frac{\alpha}{T_i^\gamma(t-1)} \log\left(\sum_{i=1}^K T_i^\gamma(t-1) \right)} \right) .$$

The idea is to 'discount' rewards that occurred far in the past, which makes the algorithm most influenced by recent events. A similar algorithm called Sliding-Window UCB uses a similar approach, but rather than discounting past rewards with a geometric discount function it simply discards them altogether. Let $\tau \in \mathbb{N}^+$ be a constant and define

$$\hat{\mu}_i^\tau(t) = \sum_{s=t-\tau+1}^{t} \mathbb{I}\{A_s = i\} X_s \qquad T_i^\tau(t) = \sum_{s=t-\tau+1}^{t} \mathbb{I}\{A_s = i\} .$$

Then the Sliding-Window UCB chooses

$$A_t = \operatorname{argmax}_{i \in [K]} \left(\hat{\mu}_i^\tau(t-1) + \sqrt{\frac{\alpha}{T_i^\tau(t-1)} \log(t \wedge \tau)} \right) .$$

It is known that if γ or τ are tuned appropriately, then for Discounted UCB the regret may be bounded by $O(\sqrt{nm}\log(n))$ and for Sliding-Window UCB by $O(\sqrt{nm\log(n)})$. Neither bound improves on what is available using Exp4, but there is some empirical evidence to support the use of these algorithms when the stochastic assumption holds.

3 An alternative way to model nonstationary stochastic bandits is to assume the mean payoffs of the arms are slowly drifting. One way to do this is to assume that $\mu_i(t)$ follows a reflected Brownian motion in some interval. It is not hard to see that the regret is necessary linear in this case because the best arm can change in any round with nonzero nondecreasing probability. The objective in this case is to understand the magnitude of the linear regret in terms of the size of the interval or volatility of the Brownian motion.

4 Yet another idea is to allow the means to change in an arbitrary way, but restrict the amount of variation. Let $\mu_t = (\mu_1(t), \ldots, \mu_K(t))$ and

$$V_n = \sum_{t=1}^{n-1} \|\mu_t - \mu_{t+1}\|_\infty$$

be the cumulative change in mean rewards measured in terms of the supremum norm. Then for each $V \in [1/K, n/K]$ there exists a policy such that for all bandits with $V_n \leq V$ it holds that

$$R_n \leq C(VK\log(K))^{1/3} n^{2/3} .$$

Furthermore, this bound is nearly tight in a minimax sense except for logarithmic terms [Besbes et al., 2014].

31.4 Bibliographic remarks

Nonstationary bandits have quite a long history. The celebrated Gittins index is based on a model where each arm is associated with a Markov chain that evolves when played and the reward depends on the state [Gittins, 1979, Gittins et al., 2011]. The classical approaches address this problem in the Bayesian framework and the objective is primarily to design efficient algorithms rather than understanding the frequentist regret. Note that the state is observed after each action. Even more related is the **restless bandit**, which is the same as Gittin's setup except the Markov chain for every action evolves in every round. The problem is made challenging because the learner still only observes the state and reward for the action they chose. Restless bandits were introduced by Whittle [1988] in the Bayesian framework and unfortunately there are more negative results than positive ones. There has been some interesting frequentist analysis, but the challenging nature of the problem makes it difficult to design efficient algorithms with meaningful regret guarantees [Ortner et al., 2012]. Certainly there is potential for more work in this area. The ideas in Section 31.1 are mostly generalizations of algorithms designed for the full information setting, notably the Fixed Share algorithm Herbster and Warmuth [1998]. The first algorithm designed for the adversarial nonstationary bandit is Exp3.S by Auer et al. [2002b]. This algorithm can be interpretted as an efficient version of Exp4 with a carefully chosen initialization such that the exponential computation over all experts collapses into a simple expression. We do not know of a clean source for this intepretation, but see the analysis of Fixed Share in the book by Cesa-Bianchi and Lugosi [2006]. The Exp3.P policy was originally developed in order to prove high probability bounds for finite-armed adversarial bandits [Auer et al., 2002b], but Audibert and Bubeck [2010b] proved that with appropriate tuning it also enjoys the same bounds as Exp3.S. Presumably this also holds for Exp3-IX. Mirror descent has been used to prove tracking bounds in the full information setting by Herbster and Warmuth [2001]. A more recent reference is by György and Szepesvári [2016], which makes the justification for clipping explicit. The latter paper considers the linear prediction setting when, similarly to the "drifting case" mentioned previously, the sequence to be compete against is unrestricted, and the goal is to guarantee that the regret scales 'nicely' with the complexity of the sequence competed against, where the complexity of a sequence is measured via the total change of subsequent vectors in the sequence. The advantage of this is that the complexity measure can distinguishes between abrupt and gradual changes. This is similar to the approach of Besbes et al. [2014] mentioned earlier. The lower bound for stochastic nonstationary bandits is by Garivier and Moulines [2011], though our proof differs in minor ways. We mentioned that there is a line of work on stochastic nonstationary bandits where the rewards are slowly drifting. The approach based on Brownian motion is due to Slivkins and Upfal [2008] while the variant described in Note 4 is by Besbes et al. [2014]. The idea of discounted UCB was introduced without analysis by Kocsis and Szepesvári

[2006]. The analysis of this algorithm and also the sliding window algorithm is by Garivier and Moulines [2011].

31.5 Exercises

31.1 Let $n, m, K \in \mathbb{N}^+$. Prove (31.2). In particular, specify first what the experts predict in each round and how Theorem 18.1 gives rise to (31.1) and how (31.2) follows from (31.1).

For the second part you may find it useful to show the following well known inequality: for $0 \leq m \leq n$, defining $\Phi_m(n) = \sum_{i=0}^{m} \binom{n}{i}$, it holds that $(m/n)^m \Phi_m(n) \leq e^m$.

31.2 Let $n, m, K \in \mathbb{N}^+$ be such that $n \geq mK$. Prove that for any policy π there exists an adversarial bandit (y_{ti}) such that

$$R_{n,m} \geq c\sqrt{nmK}\,,$$

where $c > 0$ is a universal constant.

31.3 Prove for all sufficiently large n that Exp3 from Chapter 11 has $R_{n,2} \geq cn$ for some universal constant $c > 0$.

31.4 Let $K = 2$ and $n = 1000$ and define adversarial bandit in terms of losses with $y_{t1} = \mathbb{I}\{t < n/2\}$ and $y_{t2} = \mathbb{I}\{t \geq n/2\}$. Plot the expected regret of Exp3, Exp3-IX and the variant of online stochastic mirror descent proposed in this chapter. Experiment with a number of learning rates for each algorithm.