Batched Bandits

In the standard bandit model, each reward is observed before the next arm is selected. However, in practice, this may not be the case. For example, a clinical trial may assign treatments to a group of patients before their outcomes are known. Thus, we consider the batched model, in which rewards from a batch are revealed together, after all actions have been chosen. We will consider the case of two arms and two batches, with batch sizes fixed before sampling. Following [2], we will prove a minimax regret bound and provide an algorithm that achieves it.

1 Setup

Assume that arm i has true reward θ_i. In the standard bandits setting, at each round $t \in [1, T]$, we select an arm I_t and observe $Y_t = X_{I_t,t}$, where $X_{i,t} = \theta_i + \xi_{i,t}$ and $\xi_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. We will denote $\Delta = |\theta_1 - \theta_2|$ and consider the case of $\Theta = \left\{\theta \in \mathbb{R}^2 : \Delta \leq 1\right\}$.[1] In the case of non-deterministic policies, we will let $U_1, \ldots, U_T \overset{\text{iid}}{\sim} U(0, 1)$, independently of all of the $\xi_{i,t}$, and denote $U = (U_1, \ldots, U_T)$, and let the policy be a function of U.

Definition 1 (Two-batch bandits)**.** Let $0 < t_1 < T$ be a fixed integer (possibly depending on T). In the batched setting, the policy selects actions $I_1, \ldots, I_{t_1}$ are a function of U. At round t_1, the first-batch rewards are observed, i.e. policy sees $H_{t_1} = ((I_1, Y_1), \ldots, (I_{t_1}, Y_{t_1}))$, and selects actions $I_{t_1+1}, \ldots, I_T$ as a function of (U, H_{t_1}). At round T, the second-batch rewards are observed. Let Π denote the class of all such policies, *including* all choices of t_1.

We will denote $\theta_{\max} = \max\{\theta_1, \theta_2\}$, so that the expected regret is

$$R_T(\pi, \theta) = T\theta_{\max} - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} (\theta_{\max} - \theta_{I_t})\right]$$

Definition 2. The minimax expected regret is given by

$$\mathcal{R}(T) := \inf_{\pi \in \Pi} \sup_{\theta \in \Theta} R_T(\pi, \theta)$$

i.e. the worst-case expected regret of a fixed policy π over all allowed θ.

The goal of this project is to bound the minimax expected regret:

Theorem 3. *There exist universal constants c, $C > 0$ such that, for all $T \geq 8$, we have $cT^{2/3} \leq \mathcal{R}(T) \leq CT^{2/3}$. The upper bound is achieved by Algorithm 1.*

2 Upper bound of Theorem 3

Let π be Algorithm 1, and assume WLOG that $\theta_1 \geq \theta_2$. If $\Delta = 0$ then the regret is zero. For $\Delta > 0$, each $\widehat{\theta}_i$ is the sum of $t_1/2$ independent terms each with variance 1, and thus each estimate has variance $2/t_i$. Thus, $\widehat{\theta}_1 - \widehat{\theta}_2 \sim \mathcal{N}(\Delta, 4/t_1)$. Thus, it follows from the standard Gaussian tail bound that $\mathbb{P}(\widehat{i} = 2) = \mathbb{P}\left(\frac{(\widehat{\theta}_1 - \widehat{\theta}_2) - \Delta}{\Delta\sqrt{t_1/2}}\right) \leq \exp(-t_1\Delta^2/8)$.

The first batch pulls the worse arm exactly $t_1/2$ times, incurring regret $t_1\Delta/2$. Choosing the worse arm in the second batch incurs regret $(T - t_1)\Delta$, so

$$R_T(\pi, \theta) \leq \frac{t_1\Delta}{2} + (T - t_1)\Delta \exp(-t_1\Delta^2/8)$$

[1] We impose this standard constraint as otherwise even the first round can incur arbitrarily large worst-case regret.

Algorithm 1 Explore-then-commit

Require: Known horizon $T \geq 8$.
Set $t_1 = 2\lfloor T^{2/3}/2 \rfloor$.
Pull arm 1 for $t_1/2$ rounds and arm 2 for $t_1/2$ rounds.
Observe these rewards and compute $\widehat{\theta}_i = (2/t_1)\sum_{t \leq t_1 : I_t = i} Y_t$ for $i = 1, 2$.
Choose $\widehat{i} = 1$ if $\widehat{\theta}_1 \geq \widehat{\theta}_2$, and $\widehat{i} = 2$ otherwise.
Pull arm $\widehat{i}$ for the remaining $T - t_1$ rounds.

Since $\Delta \leq 1$ and $\sup_{x \geq 0} x e^{-t_1 x^2/8} = 2e^{-1/2}/\sqrt{t_1}$, we have

$$\mathcal{R}(T) \leq R_T(\pi, \theta) \leq \frac{t_1}{2} + \frac{2e^{-1/2}(T - t_1)}{\sqrt{t_1}} \leq \frac{t_1}{2} + \frac{2e^{-1/2}T}{\sqrt{t_1}}$$

From the choice of t_1, we can select $C = \frac{1}{2} + 2\sqrt{2}e^{-1/2}$, proving the upper bound.

3 Lower bound of Theorem 3

Proposition 4. *For any $\pi \in \Pi$ and t_1,*

$$\sup_{\theta \in \Theta} R_T(\pi, \theta) \geq \max\left\{ \frac{t_1}{2}, \frac{e^{-1/2}(T - t_1)}{4\sqrt{t_1}} \right\} \tag{1}$$

The proof of Proposition 4 is dependent on the following lemma, which is proved in Perchet et al. [2, Appendix A].

Lemma 5 (Bretagnolle–Huber inequality)**.** *Let P, Q be probability distributions over the same space. For any measurable event A, $P(A) + Q(A^c) \geq \frac{1}{2}e^{-D_{KL}(P\|Q)}$.*

Proof of Proposition 4. For the first term, the supremum is bounded below by

$$\begin{aligned}
\max_{\theta \in \{(1,0),(0,1)\}} R_T(\pi, \theta) &\geq \frac{1}{2}R_T(\pi, (1,0)) + \frac{1}{2}R_T(\pi, (0,1)) \\
&\geq \frac{1}{2}R_{t_1}(\pi, (1,0)) + \frac{1}{2}R_{t_1}(\pi, (0,1)) \\
&= \frac{1}{2}|\{t \leq t_1 : I_t = 2\}| + \frac{1}{2}|\{t \leq t_1 : I_t = 1\}| = \frac{t_1}{2}
\end{aligned}$$

For the second term, similarly the supremum is similarly bounded below by the average regret of $\theta \in \{(\delta, 0), (0, \delta)\}$ where $\delta = t_1^{-1/2} \leq 1$. Note that each second-batch choice depends only on $Z := (U, I_1, Y_1, \ldots, I_{t_1}, Y_{t_1})$, i.e. for $t > t_1$, we may write $I_t = f_t(Z)$ where f_t is deterministic. Let P denote the distribution of Z under $\theta = (\delta, 0)$ and Q its distribution under $\theta = (0, \delta)$. It follows that

$$\begin{aligned}
\sup_{\theta \in \Theta} R_T(\pi, \theta) &\geq \frac{1}{2}R_T(\pi, (\delta, 0)) + \frac{1}{2}R_T(\pi, (0, \delta)) \\
&\geq \frac{1}{2}R_{t_1+1:T}(\pi, (\delta, 0)) + \frac{1}{2}R_{t_1+1:T}(\pi, (0, \delta)) \\
&= \frac{\delta}{2}\sum_{t=t_1+1}^{T}\left[P(f_t(Z) = 2) + Q(f_t(Z) = 1)\right] \\
&\geq \frac{\delta(T - t_1)}{4}e^{-D_{\mathrm{KL}}(P\|Q)}
\end{aligned} \tag{2}$$

where in the last line we have used Lemma 5. Now, we will compute $D_{\mathrm{KL}}(p \parallel q)$. Let $p_\mu(y) = (2\pi)^{-1/2} e^{-(y-\mu)^2/2}$ be the density of $N(\mu, 1)$. For $Y \sim N(a, 1)$,

$$\mathbb{E}\left[\log \frac{p_a(Y)}{p_b(Y)}\right] = \frac{1}{2}\mathbb{E}\left[(Y-b)^2 - (Y-a)^2\right] = \frac{1}{2}(a-b)^2$$

implies that

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \mathbb{E}_{Z\sim P}\left[\sum_{t\le t_1 : I_t = 1} \log \frac{p_\delta(Y_t)}{p_0(Y_t)} + \sum_{t \le t_1 : I_t = 2} \log \frac{p_0(Y_t)}{p_\delta(Y_t)}\right] \\ &= \frac{1}{2}\delta^2 |\{t \le t_1 : I_t = 1\}| + \frac{1}{2}\delta^2 |\{t \le t_1 : I_t = 2\}| = \frac{1}{2} t_1 \delta^2 = \frac{1}{2} \end{aligned}$$

where in the first line we have used that $(U, I_1, \ldots, I_{t_1})$ have the same distribution under P and Q. Substituting into Equation (2) completes the proof. ■

Lower bound of Theorem 3. If $t_1 \ge T^{2/3}$, the first term of Equation (1) is at least $T^{2/3}/2$. If $t_1 < T^{2/3}$, then $T - t_1 \ge T/2$ for all $T \ge 8$ and $\sqrt{t_1} < T^{1/3}$, and the second term of Equation (1) is at least $e^{-1/2} T^{2/3}/8$. Thus, we can select $c = e^{-1/2}/8$ in either case. ■

4 Extensions

Definition 1 can be easily extended to $M \ge 2$ batches, via $0 = t_0 < t_1 < \cdots < t_M = t$, and Perchet et al. [2, Theorems 3–4] shows that our $M = 2$ example is just a special case. For example, the $M = T$ batch case reduces to sequential bandits. Gao et al. [1] extends this to $n \ge 2$ arms under the assumption that $\Delta_i = \max_j \theta_j - \theta_i \le \sqrt{n}$. Gao et al. [1] also considers the setting where the boundaries are adaptive, and finds that the regret bounds have the same powers of T up to logarithmic factors.

Shu et al. [3] extends this to the case where the rewards may be heavy-tailed instead of Gaussians. Specifically, they show that in the case where $\mathbb{E}|\xi_{i,t}|^{1+\varepsilon} \le B$ for known $B > 0$ and $\varepsilon \in (0, 1]$. In the two-arm, two-batch setting, they prove minimax regret bounds of order $T^{(1+\varepsilon)/(1+2\varepsilon)}$ up to log factors. Our analysis considered the special case $\varepsilon = 1$, which recovers order $T^{2/3}$.

One future direction is as follows: the problem setting is restricted to fixed finite horizon T. A more general setting may consider a termination time $\tau = \inf\{t_i : s(H_{t_i}) = 1\}$ for a given stopping rule s. In this setting, our goal is to minimize discounted regret $\mathbb{E}\left[\sum_{t=1}^{\tau} \gamma^{t-1}(\theta_{\max} - \theta_{I_t})\right]$. Another is that θ may evolve over time, and thus the selection of $t_1, \ldots, t_m$ becomes a part of the policy.

References

[1] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. 40:1456–1456, 03–06 Jul 2015. URL https://proceedings.mlr.press/v40/Perchet15.html.

[3] Yunlu Shu, Yunwen Guo, Gongyi Zhuo, and Tianyu Wang. Batched bandits with heavy-tailed rewards. *Information and Inference: A Journal of the IMA*, 15(3):iaag031, 09 2026. ISSN 2049-8772. doi: 10.1093/imaiai/iaag031. URL https://doi.org/10.1093/imaiai/iaag031.

AI statement. GPT-6.1 was used to check derivations and clean up notation.