A PDE-Based Analysis of the Symmetric Two-Armed Bernoulli Bandit

Vlad Kobzar[1]
joint work with Robert Kohn[2]

[1]APAM, Columbia

[2]Courant Institute, NYU

Sept 13, 2023

Symmetric Two-Armed Bernoulli Bandit

- Background
- Minimax optimal player
- PDE-based characterization of regret

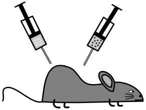
Slot machine

Round	1	2	3	4	5	6	7	8	9	10	Total
Left (arm 1)	0		1	0		0				1	$G_1 = 2$
Right (arm 2)		1			1		1	0	0		$G_2 = 3$

*The distribution of the arms is not directly revealed to the player

Image: (LS20)

A/B testing

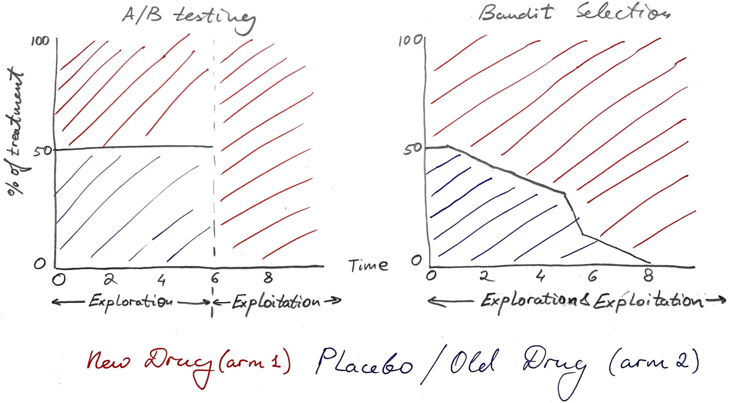


- Let s_i be the sample size for i-th 'arm' (drug 1 vs placebo 2).
$$T = s_1 + s_2$$
- For the expected probability of recovery m_i,
$$H_0 = \{m_1 \leq m_2\}$$
- A test statistic
$$z = \left(\bar{m}_1 - \bar{m}_2\right) \Big/ \sqrt{\bar{\sigma}_1^2/s_1 + \bar{\sigma}_2^2/s_2}$$
where the sample mean & variance for Bernoulli r.v.'s are resp.
$$\bar{m}_i = G_i/s_i \text{ and } \bar{\sigma}_i^2 = \bar{m}_i(1 - \bar{m}_i)$$
- Reject H_0 if, e.g.,
$$z > \Phi^{-1}(.95) = 1.645$$

Image: (LS20)

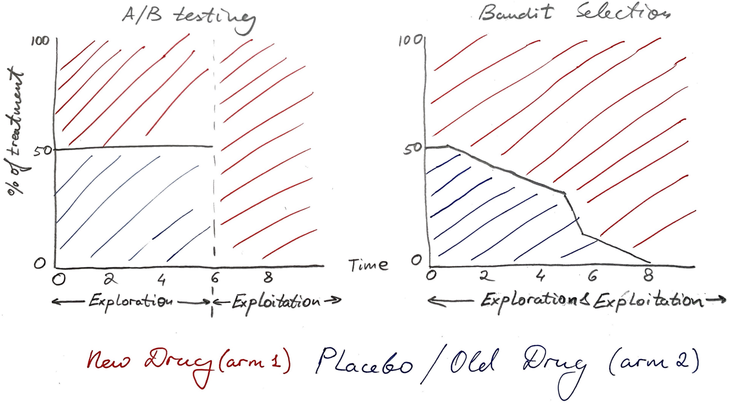
Thompson, 1933

- The data so far suggests that the new drug is better, but is not "conclusive" (as defined by $p - \text{value} < 0.05$)
- Should we adjust our actions to minimize the administration of the inferior treatment?
- If 90% confident, allocate 90% of the trials to the new drug



Exploration/exploitation trade-off

- Exploitation: choose the best action given the revealed data
- Exploration: choose the best action to improve our knowledge about the values of different actions
- Bandit selection entails a trade-offs between these extremes



Two-armed bandit

- One of the oldest learning problems (Tho33; Rob52)
- The distributions of the arms $a := (a_1, a_2)$ is secret
- The player selects the arm according to a policy $(p_t)_{t\in[T]}$
- Each p_t is based on information revealed earlier, at each $\tau < t$

In each $t \in [T]$

1. $g_t \sim a$
2. $I_t \sim p_t$ (independently of g_t)
3. I_t and $g_{I_t,t}$ are revealed to the player

Pseudoregret

- The *gap* between the arms' means:

$$\epsilon = m_1 - m_2$$

- The final-time expected *pseudoregret*:

$$\bar{R}_T(p, a) := \epsilon \mathbb{E}_{p,a} s_2$$

  where s_2 = sample size for 2nd (suboptimal) arm
- The expected loss from choosing the suboptimal arm
- Drop the subscript in s_2 for simplicity going forward

Regret

$\mathbb{E}$(rewards of the best arm in hindsight)$-$(rewards of the chosen arms)

- Formalized as

$$R_T(p,a) = \mathbb{E}_{p,a} \max_{k\in\{1,2\}} \sum_t (g_{k,t} - g_{I_t,t})$$

- Equivalently

$$R_T(p,a) := \mathbb{E}_{p,a} \max_{k\in\{1,2\}} x_{k,0}$$

where

$$x_t := \sum_{\tau<t} g_\tau - g_{I_\tau,\tau}\mathbb{1}$$

- (Denote the starting time by -T and final time by 0)

Pseudoregret vs regret

- We have

$$\bar{R}_T(p,a) \le R_T(p,a)$$

- From the previous slide

$$R_T(p,a) := \mathbb{E}_{p,a} \max_{k\in\{1,2\}} x_{k,0}$$

- Also

$$\begin{aligned}
\bar{R}_T(p,a) &:= \epsilon \mathbb{E}_{p,a} s \\
&= \mathbb{E}_{p,a} \sum_t (m_1 - m_2) \mathbb{1}_{I_t=2} \\
&= \max_{k\in\{1,2\}} \mathbb{E}_a \sum_t g_{k,t} - \mathbb{E}_{p,a} \sum_t g_{I_t,t} \\
&= \max_{k\in\{1,2\}} \mathbb{E}_{p,a} x_{k,0}
\end{aligned}$$

Minimax (pseudo) regret

- The *minimax* regret

$$R_T^* := \min_p \max_a R_T(p, a) \tag{1}$$

- Similarly, the *minimax* pseudoregret

$$\bar{R}_T^* := \min_p \max_a \bar{R}_T(p, a) \tag{2}$$

- Even when the arms are iid Bernoulli, these quantities (or any *minimax optimal player* that attains them) are not known exactly.

Bernoulli two-armed bandits - fixed gap

- Fixed gap regime

$$\min_{p} \bar{R}_T(a, p) = O\Big(\frac{1}{\epsilon}\log T\Big)$$

- The bound is vacuous when

$$\epsilon \to 0$$

  fast enough
- I.e., the difficulty of detecting the gap increases as the sample size increases

Bernouilli two-armed bandits - minimax setting

- Asymptotic bounds (Bat83; Vog60)

$$0.306 \leq \lim_{T\to\infty} \inf \bar{R}^*_T/\sqrt{T} \leq \lim_{T\to\infty} \sup \bar{R}^*_T/\sqrt{T} \leq 0.376$$

- Nonasymptotic bounds (BCB12; RVR14; LG21)

$$0.07\sqrt{T} \leq \begin{cases} \bar{R}^*(T) \leq .832\sqrt{T} \\ R^*(T) \leq 7.762\sqrt{T} \end{cases}$$

- The exact constants in front of $\sqrt{T}$ are still unknown.
- The nonasymptotic l.b. is achieved by the *symmetric* Bernoulli bandit
  - The only known lower bound for k-armed bandits too

Symmetric two-armed Bernoulli bandit

- Optimal arm a_1: for $0 < \epsilon < 1$,

$$P(a_1 = -1) = \frac{1-\epsilon}{2} \quad \text{and} \quad P(a_1 = 1) = \frac{1+\epsilon}{2}$$

- Suboptimal arm a_2:

$$P(a_i = -1) = \frac{1+\epsilon}{2} \quad \text{and} \quad P(a_i = 1) = \frac{1-\epsilon}{2}$$

  independently from arm 1 and the history
- The arms are *statistically equivalent*

$$a_1 = -a_2$$

- I.e., samples from one arm can be converted into samples from the other one by flipping the sign

Myopic player

- In general good players balance sampling of the arms (*exploration*) with choosing the arm with the highest expected reward (*exploitation*).
- But in the symmetric case, the distribution of arms 1 and 2 are *statistically equivalent*

$$a_1 = -a_2$$

- Thus, the player will get the same information about the means of both arms by sampling either one
- While exploration is not needed, this symmetric problem has not been fully understood previously

Previous results - symmetric case

- (Bat83) considered a *Bayesian* version of the symmetric two-armed bandit problem
- Same as above except that the index of the optimal arm j is sampled from a uniform distribution over $\{1, 2\}$
- The Bayes optimal player is myopic = chooses the arm with the highest posterior probability of being the safe one given the revealed rewards.

Previous results - symmetric case

- The Bayes optimal (myopic) policy can be computed explicitly.

$$p^m = \begin{cases} (1,0) \text{ if } G_1 - G_2 > 0 \\ \left(\frac{1}{2}, \frac{1}{2}\right) \text{ if } G_1 - G_2 = 0 \\ (0,1) \text{ if } G_1 - G_2 < 0 \end{cases}$$

where the revealed *cumulative rewards* of arm i are

$$G_i = \sum_{\tau < t} g_{i,\tau} \mathbb{1}_{I_\tau = i},$$

- (Bat83) determined the leading order of *Bayesian* pseudoregret under the worst case (uniform) prior

$$\bar{R}_T^B \approx .530\sqrt{T}$$

Our advances - symmetric two-armed bandit

- Prove the *minmax* optimality of p^m
- Associate the leading order term of R_T^* and $\bar{R}_T^*$ with explicit solutions of linear heat equations
- Make progress towards unifying the analyses of
  - Bayesian and minimax regret, on the one hand, and
  - regret and pseudoregret, on the other hand.

Main ideas - minimax optimality of p^m

- Centering the arms helps simplify the state space (eliminate dependence of s_i)
- As noted an optimal strategy can depend on all available information at time $t > -T$, i.e., $p_t \equiv p_t(H_{t-1})$ where

$$H_{t-1} := (I_{-T}, ..., I_{t-1}, g_{I_{-T},-T}, ..., g_{I_{t-1},t-1}), \tag{3}$$

- Unlike the Bayesian setting where the posterior depends on cumulative statistics, in the minimax setting we need to show that the minimax optimal player depends on the cumulative and not entire history
- We show that if the optimal player depends on the full history, it can be modified to depend on cumulative statistics only while preserving optimality

Main ideas - PDE characterization of (pseudo)regret

- "Numerical analysis in reverse"
- Characterize the regret and pseudoregret by dynamic programs
- Find the PDEs that are discretized by these DPs

State variables

- The centered revealed rewards are

$$\xi := \sum_{\tau < t} \left(g_{1,\tau} \mathbb{1}_{I_\tau = 1} - g_{2,\tau} \mathbb{1}_{I_\tau = 2} \right) - \epsilon t$$

- The final pseudoregret is proportional to the number of times the suboptimal arm was sampled

$$2\epsilon s$$

- $-T$ is the starting time and 0 is the final time

Discrete iterative scheme

- Expected final-time regret achieved by the player p if the prediction process starts at time t w/given ξ and s:

$$v(\xi, s, 0) = 2\epsilon s \tag{4a}$$
$$v(\xi, s, t) = \mathbb{E}_{a,p}\ v(\xi + d\xi, s + ds, t + 1) \text{ for } t \leq -1 \tag{4b}$$

where

$$ds = \mathbb{1}_{I=2} \quad \text{and} \quad d\xi = g_1 \mathbb{1}_{I=1} - g_2 \mathbb{1}_{I=2} - \epsilon$$

- The PDE comes from the heuristics that the definition of v should be a consistent numerical scheme for the PDE

PDE associated with the discrete scheme

Let u be the C^0 solution on $\mathbb{R}^3 \times \mathbb{R}_{<0}$ of

$$u_t + Lu = q \tag{5a}$$

$$u(\xi, s, 0) = 2\epsilon s \tag{5b}$$

where

$$Lu = \frac{\kappa}{2} u_{\xi\xi} \text{ and } \kappa = 1 - \epsilon^2,$$

and the stationary source is

$$q(\xi + \epsilon t) = \begin{cases} -2\epsilon & \text{if } \xi + \epsilon t < 0 \\ 0 & \text{if } \xi + \epsilon t > 0 \end{cases}$$

- L is obtained by computing the leading terms of the Taylor expansion of $\mathbb{E}_{a,p}\, v(\xi + d\xi, s + ds, t + 1)$ around $v(\xi, s, t)$
- q is obtained by comparing $\mathbb{E}_{a,p}\, v(\xi, s + ds, t)$ with $v(\xi, s, t)$ assuming $v(\xi, s + c, t) = v(\xi, s, t) + \epsilon c$ like the final value
- $u_{\xi\xi}$ (and possibly u_ξ) are discontinuous at $\xi = 0$ due to the discontinuity of q.

The PDE has a explicit solution

- The fundamental solution G of $w_t + Lw = 0$ is explicit
- Let φ be a C^0 function of $y = \xi + \epsilon t$ satisfying

$$L\varphi = \epsilon\varphi' + \frac{\kappa}{2}\varphi'' = q \tag{6}$$

- It is explicit, and we can choose constants to ensure at most linear growth at ∞

$$\varphi(y) = \begin{cases} -2y & \text{if } y \leq 0 \\ be^{-2\epsilon y} - b & \text{if } y > 0 \end{cases}$$

  where b parametrizes the jump of φ' at $\xi + \epsilon t = 0$.
- Then for $w = u - \varphi$,

$$w_t + Lw = 0$$
$$w(\xi, s, 0) = \psi(\xi, s)$$

  where $\psi(\xi, s) = 2\epsilon s - \varphi(\xi)$.
- Therefore, u is also explicit:

$$u = G * \psi + \varphi = 2\epsilon s - G * \varphi + \varphi$$

Backwards induction (verification argument)

- Initialization $u(\xi, s, 0) = v(\xi, s, 0)$
- Hypothesis $v(\xi, s, t+1) = u(\xi, s, t+1) + E(t+1)$
- The discretization error E is bounded by bounding the higher order terms of the Taylor expansion of u

$$
\begin{aligned}
& u(\xi, s, t) + E(t) \\
& = \mathbb{E}_{p,a}\ u(\xi + d\xi, s + ds, t+1) + E(t+1) && \textbf{[PDE+higher term bound]} \\
& = \mathbb{E}_{p,a}\ v(\xi + d\xi, s + ds, t+1) && \textbf{[inductive hypothesis]} \\
& = v(\xi, s, t) && \textbf{[DPP of } v\textbf{]}
\end{aligned}
$$

Discretization error

- How is Taylor expansion useful in a game where the rewards are not small?
- As noted before

  $$u = 2\epsilon s - G * \varphi + \varphi$$

  where G is a fundamental solution of a linear fixed coefficients parabolic PDE $w_t + Lw = 0$
- G is a function of $\xi/\sqrt{|t|}$. Therefore, when $|t|$ is large, $\xi + d\xi$ is only $O(1/\sqrt{|t|})$ apart from ξ
- 3rd spatial derivatives of $G * \varphi$ are $O(|t|^{-\frac{3}{2}})$ –integrating them with respect to t leads to an $O(1)$ cumulative error
- Also $\varphi^{(d)} = O(b\epsilon^d)$, which is controlled by the rate at which $b\epsilon \to 0$ approaches zero as $T \to \infty$

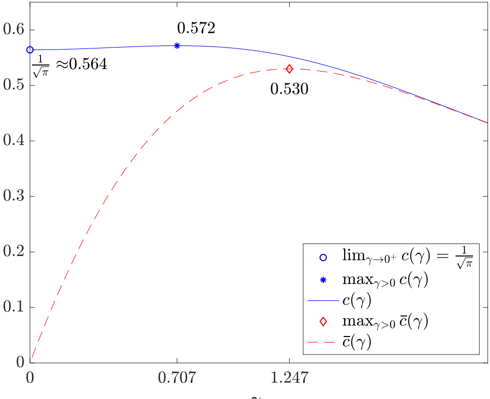
Medium gap

- Let $\epsilon = \gamma/\sqrt{T}$ for constant $\gamma > 0$. Then

$$c(\gamma) = \lim_{T\to\infty} \frac{1}{\sqrt{T}} R_T(p^m, a) = \lim_{T\to\infty} \frac{1}{\sqrt{T}} u(0,0,-T) =$$

$$= \frac{1}{\sqrt{\pi}} e^{-\gamma^2} + \gamma \mathrm{erf}(\gamma) + \left(\frac{1}{\gamma} - \gamma\right) \mathrm{erf}\left(\frac{\gamma}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} e^{-\frac{\gamma^2}{2}}$$

- Similarly compute the prefactor $\bar{c}$ of the leading order of $\bar{R}_T$



Other gap regimes

- We can also compute $R_T(p^m, a)$ and $\bar{R}_T(p^m, a)$ when ϵ approaches zero faster and slower than $\gamma/\sqrt{T}$

	Small gap $\epsilon = o\left(T^{-\frac{1}{2}}\right)$	*Medium gap* $\epsilon = \gamma T^{-\frac{1}{2}}$	*Large gap:* $\epsilon \in \left[\omega\left(T^{-\frac{1}{2}}\right), o(1)\right]$
$R_T(p^m, a)$ $\min(E_1(-T), E_0(-T))$	$\frac{1}{\pi} T^{\frac{1}{2}} \approx .564\ T^{\frac{1}{2}}$ $O(1)$	$c(\gamma) T^{\frac{1}{2}} (\max .572\ T^{\frac{1}{2}})$ $O(1)$	$1/\epsilon$ $O(1)$
$\bar{R}_T(p^m, a)$ $\min(\bar{E}_1(-T), \bar{E}_0(-T))$	ϵT $O(\epsilon \log T + \epsilon^2 T)$	$\bar{c}(\gamma) T^{\frac{1}{2}} (\max .530\ T^{\frac{1}{2}})$ $O(1)$	$1/\epsilon$ $O(1)$

Our advances - symmetric two-armed bandit

- Since p^m is discontinuous as a function of revealed gains when $G_1 - G_2 = 0$, the spatial derivatives of the solution of the relevant PDEs are discontinuous in that region.
  - While this discontinuity does not affect the leading order behavior, it affects the discretization error.
  - We optimize this discontinuity to minimize this error.
- We explicitly determine the leading order terms of R_T^* and $\bar{R}_T^*$, i.e., the exact constants in front of $\sqrt{T}$, in the symmetric bandit setting.
  - Existing techniques rely on information theory to bound below $\bar{R}_T^*$ in the Bernouilli bandit problems.
  - This leads to the only known nonasymptotic lower bounds in the general bandit problems.
  - Our results lead to sharper nonasymptotic regret and pseudoregret lower bounds in the 2-armed case
  - They are established by more elementary methods.

Extensions

- Determining optimal policies in general bandits is more challenging due to the exploration- exploitation trade-off
- But there are more realistic settings where exploration is not needed, and therefore we expect that our work can be extended, such as
  - Bayesian k-armed bandit setting where the player knows that one arm has an arbitrary distribution P (but does not know which arm) while all the other arms have the same distribution Q (Fel62; Rod78; Zab76).
  - The only known lower bound for k-armed bandits is established using such a distribution
- General bandit problems where a (possibly suboptimal) policy is given (e.g., Thompson sampling)

Conclusion

- Offer a fresh PDE-based perspective on symmetric two-armed bandit problem
- Provide minmax optimality of the myopic player and obtain the leading order minimax optimal regret and pseudoregret with explicit non-asymptotic error bounds
- This improves the corresponding lower bounds for the general two-armed Bernoulli bandit
- Develop novel PDE-based techniques that could be used to investigate bandit problems further
- For more details see our preprint (KK22) on arXiv

Acknowledgements

- V.A.K is supported by NSF grant DMS-1937254. R.V.K. is supported by NSF grant DMS-2009746.

References I

[Bat83] J. A. Bather, *The minimax risk for the two-armed bandit problem*, Mathematical Learning Models — Theory and Algorithms (New York, NY) (Ulrich Herkenrath, Dieter Kalin, and Walter Vogel, eds.), Springer New York, 1983, pp. 1–11.

[BCB12] Sébastien Bubeck and Nicolò Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Foundations and Trends® in Machine Learning **5** (2012), no. 1, 1–122.

[Fel62] Dorian Feldman, *Contributions to the 'two-armed bandit' problem*, Ann. Math. Statist. **33** (1962), 847–856.

[KK22] Vladimir A. Kobzar and Robert V. Kohn, *A PDE-based analysis of the symmetric two-armed bernoulli bandit*, 2022.

References II

[LG21] Tor Lattimore and Andras Gyorgy, *Mirror descent and the information ratio*, Proceedings of Thirty Fourth Conference on Learning Theory (Mikhail Belkin and Samory Kpotufe, eds.), Proceedings of Machine Learning Research, vol. 134, PMLR, 15–19 Aug 2021, pp. 2965–2992.

[LS20] Tor Lattimore and Csaba Szepesvari, *Bandit algorithms*, Cambridge University Press, 2020.

[Rob52] Herbert Robbins, *Some aspects of the sequential design of experiments*, Bulletin of the American Mathematical Society **58 (5)** (1952), 527–535.

[Rod78] Leiba Rodman, *On the many-armed bandit problem*, Ann. Probab. **6(3)** (1978), 491–498.

References III

[RVR14] Daniel Russo and Benjamin Van Roy, *Learning to optimize via information-directed sampling*, Advances in Neural Information Processing Systems (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[Tho33] William R. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika **25(3–4)** (1933), 285–294.

[Vog60] Walter Vogel, *An asymptotic minimax theorem for the two armed bandit problem*, Ann. Math. Statist. **31** (1960), no. 2, 444–451.

[Zab76] A. A. Zaborskis, *Sequential bayesian plan for choosing the best method of medical treatment.*, Avtomatika i Telemekhanika **11** (1976), 144–153.