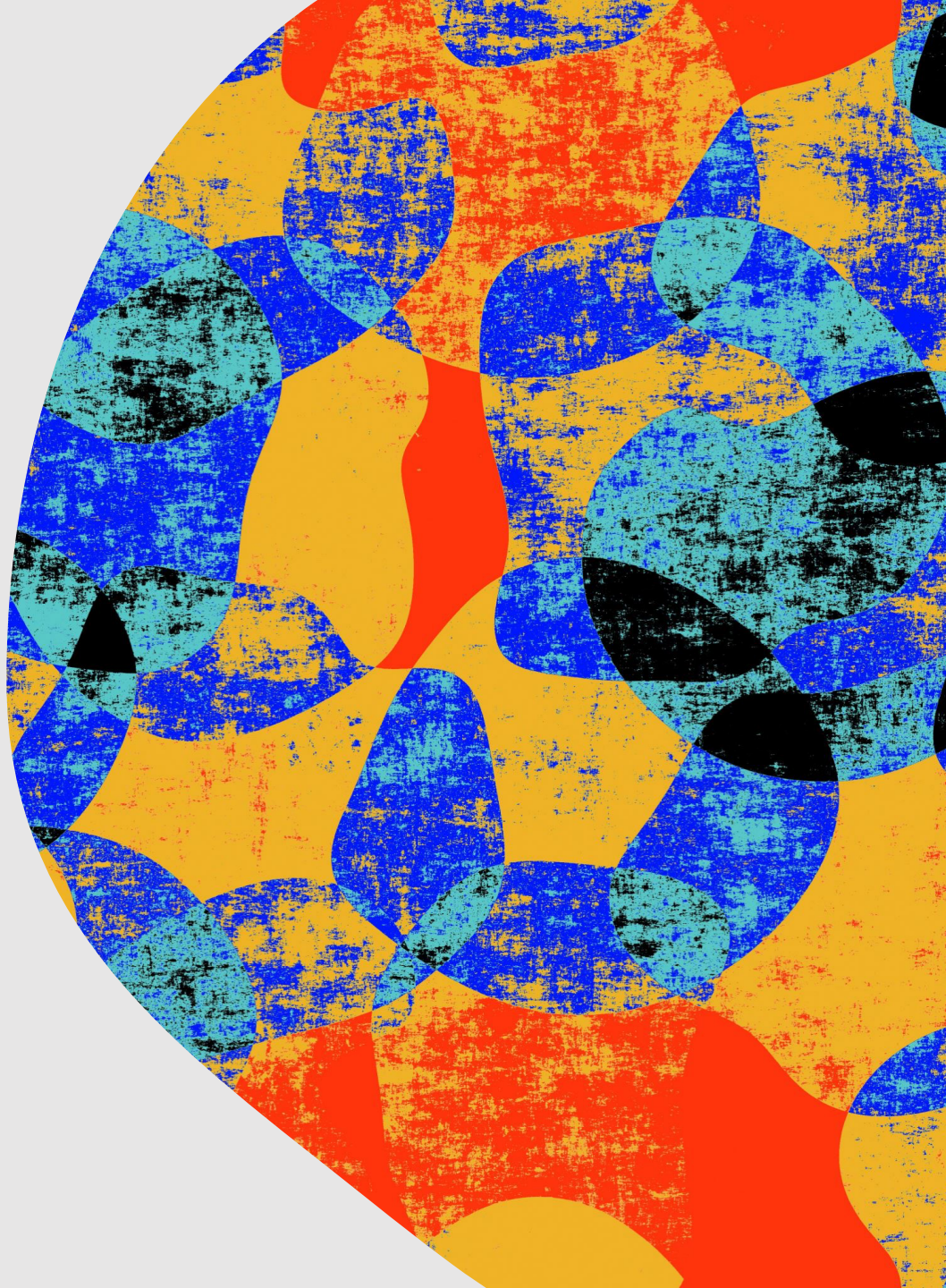
Bandit Algorithms Chapter 4: Stochastic Bandits

Authors: Tor Lattimore and Csaba Szepesvári

Slides: Hunter Heidenreich



Core Assumptions

$$\nu = (P_a : a \in \mathcal{A})$$

Core Assumptions

Distribution as indexed by selected action

Set of actions

$$\nu = (P_a : a \in \mathcal{A})$$

A stochastic bandit

Core Assumptions

- A learner interacts with the bandit:
  - Time horizon: n
  - Specific time step: t (in {1, …, n})
  - Specific action selected: A_t
  - Specific reward received: X_t
- This creates a *probability measure* over the sequence:
  - $A_1,X_1,A_2,X_2,\ldots,A_n,X_n$

Core Assumptions – Sequence Assumptions

$$p(X_t | A_1, X_1, \ldots, A_{t-1}, X_{t-1}, A_t) = P_{A_t}$$

dictated by

$$\pi_t(\cdot | A_1, X_1, A_2, X_2, \ldots, A_{t-1}, X_{t-1})$$

No access to the future

The Learning Objective

$$S_n = \sum_{t=1}^{n} X_t$$

(reward maximization)

??? Why is this not optimization??

- Unknown time horizon
  - (sometimes we know, sometimes we don't)
  - (not a serious issue and is typically easy to handle)
- Cumulative reward is a random quantity
- Unknown distributions for each arm

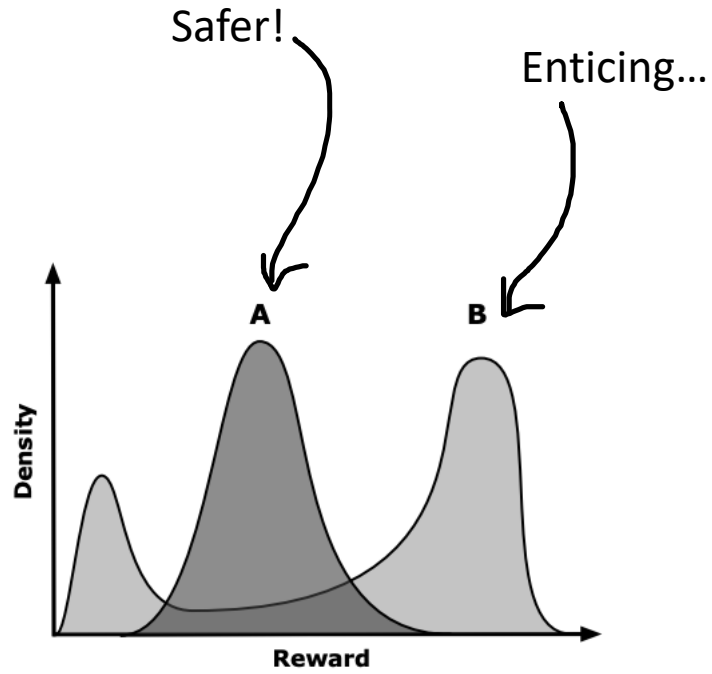


Figure 4.1 Alternative revenue distributions

Environmental Class

What do we know about the class of problems our learner will interact with?

Unstructured

- No action produces information about another action
- More formally:

$$\mathcal{E} = \{\nu = (P_a : a \in \mathcal{A}) : P_a \in \mathcal{M}_a, \forall a \in \mathcal{A}\}$$

Name	Symbol	Definition
Bernoulli	$\mathcal{E}_{\mathcal{B}}^k$	$\{(\mathcal{B}(\mu_i))_i : \mu \in [0,1]^k\}$
Uniform	$\mathcal{E}_{\mathcal{U}}^k$	$\{(\mathcal{U}(a_i,b_i))_i : a,b \in \mathbb{R}^k \text{ with } a_i \leq b_i \text{ for all } i\}$
Gaussian (known var.)	$\mathcal{E}_{\mathcal{N}}^k(\sigma^2)$	$\{(\mathcal{N}(\mu_i,\sigma^2))_i : \mu \in \mathbb{R}^k\}$
Gaussian (unknown var.)	$\mathcal{E}_{\mathcal{N}}^k$	$\{(\mathcal{N}(\mu_i,\sigma_i^2))_i : \mu \in \mathbb{R}^k \text{ and } \sigma^2 \in [0,\infty)^k\}$
Finite variance	$\mathcal{E}_{\mathbb{V}}^k(\sigma^2)$	$\{(P_i)_i : \mathbb{V}_{X\sim P_i}[X] \leq \sigma^2 \text{ for all } i\}$
Finite kurtosis	$\mathcal{E}_{\text{Kurt}}^k(\kappa)$	$\{(P_i)_i : \text{Kurt}_{X\sim P_i}[X] \leq \kappa \text{ for all } i\}$
Bounded support	$\mathcal{E}_{[a,b]}^k$	$\{(P_i)_i : \text{Supp}(P_i) \subseteq [a,b]\}$
Subgaussian	$\mathcal{E}_{\text{SG}}^k(\sigma^2)$	$\{(P_i)_i : P_i \text{ is } \sigma\text{-subgaussian for all } i\}$

Table 4.1 Typical environment classes for stochastic bandits. $\text{Supp}(P)$ is the (topological support of distribution P. The kurtosis of a random variable X is a measure of its tai behaviour and is defined by $\mathbb{E}[(X-\mathbb{E}[X])^4]/\mathbb{V}[X]^2$. Subgaussian distributions have simila properties to the Gaussian and will be defined in Chapter 5.

Examples of Unstructured Environment Classes

Structured Bandits

- Actions are correlated in some manner
- Any bandit problem that is not *strictly unstructured*
- A learner can obtain info about an action it never plays!

EXAMPLE 4.1. Let $\mathcal{A} = \{1, 2\}$ and $\mathcal{E} = \{(\mathcal{B}(\theta), \mathcal{B}(1-\theta)) : \theta \in [0, 1]\}$. In this environment class, the learner does not know the mean of either arm, but can learn the mean of both arms by playing just one. The knowledge of this structure dramatically changes the difficulty of learning in this problem.

Learner Performance – The Regret

Defined for:

- A specific learner policy
- A specific bandit instance
- A specific time horizon

The largest mean of all actions, defined by the instance

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

The Regret Lemma

LEMMA 4.4. *Let ν be a stochastic bandit environment. Then,*

(a) $R_n(\pi, \nu) \geq 0$ *for all policies π;*

- No negative regret
- It is impossible to better than selecting an optimal action for every round

The Regret Lemma

LEMMA 4.4. *Let ν be a stochastic bandit environment. Then,*

(a) $R_n(\pi, \nu) \geq 0$ *for all policies* π*;*
(b) *the policy* π *choosing* $A_t \in \operatorname{argmax}_a \mu_a$ *for all* t *satisfies* $R_n(\pi, \nu) = 0$*;*

- There exists a policy that achieves 0 regret
- That policy is the policy of selecting an action in the optimal set

The Regret Lemma

LEMMA 4.4. *Let ν be a stochastic bandit environment. Then,*

(a) $R_n(\pi, \nu) \geq 0$ *for all policies* π*;*
(b) *the policy* π *choosing* $A_t \in \operatorname{argmax}_a \mu_a$ *for all* t *satisfies* $R_n(\pi, \nu) = 0$*; and*
(c) *if* $R_n(\pi, \nu) = 0$ *for some policy* π*, then* $\mathbb{P}(\mu_{A_t} = \mu^*) = 1$ *for all* $t \in [n]$.

- Achieving 0 regret is possible **if and only if** the learner knows the exact bandit that is being faced beforehand

Regret Objectives – Sub-linear

$$\text{for all } \nu \in \mathcal{E}\,, \qquad \lim_{n\to\infty} \frac{R_n(\pi,\nu)}{n} = 0\,.$$

$$\text{for all } \nu \in \mathcal{E}\,, \qquad R_n(\pi,\nu) \leq Cn^p.$$

p < 1

Regret Objectives – Factorization

$$\text{for all } n \in \mathbb{N},\ \nu \in \mathcal{E}\,, \qquad R_n(\pi, \nu) \leq C(\nu) f(n)\,.$$

Regret due to horizon

Regret due to instance

Decomposing the Regret

LEMMA 4.5 (Regret decomposition lemma). *For any policy π and stochastic bandit environment ν with $\mathcal{A}$ finite or countable and horizon $n \in \mathbb{N}$, the regret R_n of policy π in ν satisfies*

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}\left[T_a(n)\right] . \tag{4.5}$$

$$\Delta_a(\nu) = \mu^*(\nu) - \mu_a(\nu)$$

$$T_a(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$$

Bernoulli Bandit and Follow-The-Leader

Follow-The-Leader Analysis

- 2-Arm Bernoulli Bandit
  - 1 arm with mean = 0.5
  - 1 arm with mean = 0.6
  - Horizon (n = 100 rounds)
- What's the best that could happen?

Follow-The-Leader Analysis

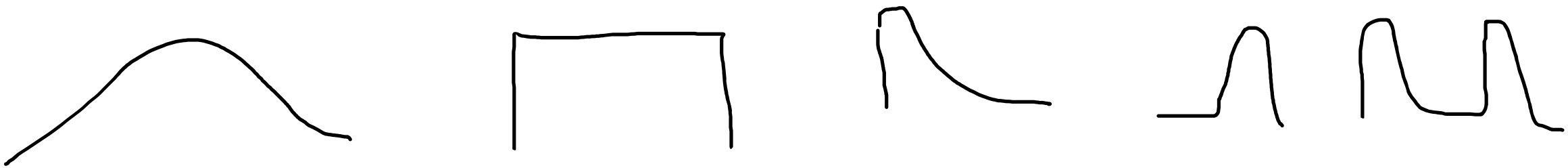
- 2-Arm Bernoulli Bandit
  - 1 arm with mean = 0.5
  - 1 arm with mean = 0.6
  - Horizon (n = 100 rounds)
- What's the best that could happen?
  - Must suffer 0.1 regret (by sampling each arm once)
  - Could select 0.6 arm the rest of the time
  - Min bound: 0.1!
- What's the worst?

Follow-The-Leader Analysis

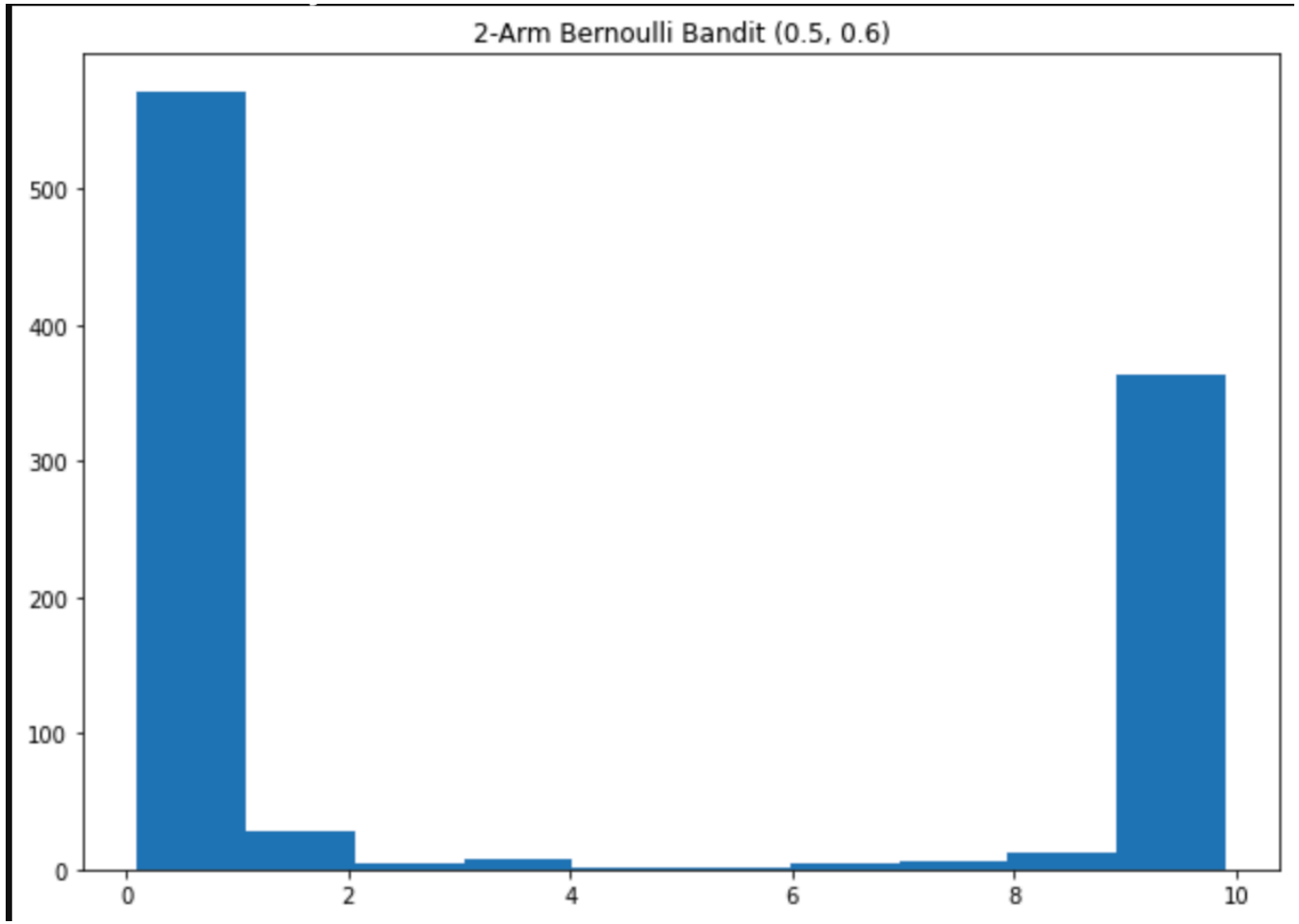
- 2-Arm Bernoulli Bandit
  - 1 arm with mean = 0.5
  - 1 arm with mean = 0.6
  - Horizon (n = 100 rounds)
- What's the best that could happen?
  - Must suffer 0.1 regret (by sampling each arm once)
  - Could select 0.6 arm the rest of the time
  - Min bound: 0.1!
- What's the worst?
  - 99 * 0.1 = 9.9!

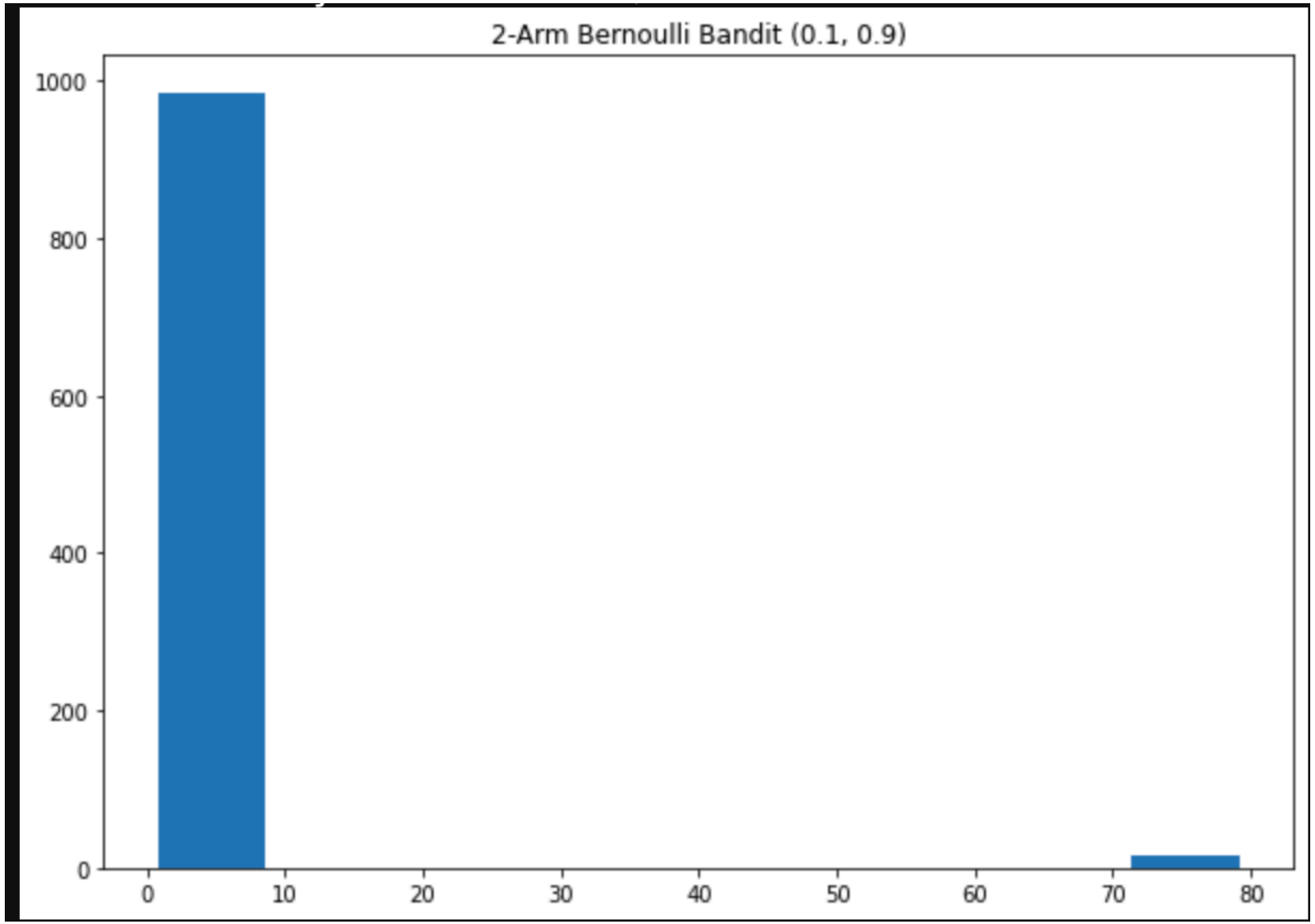
Simulating this experiment 1000 times

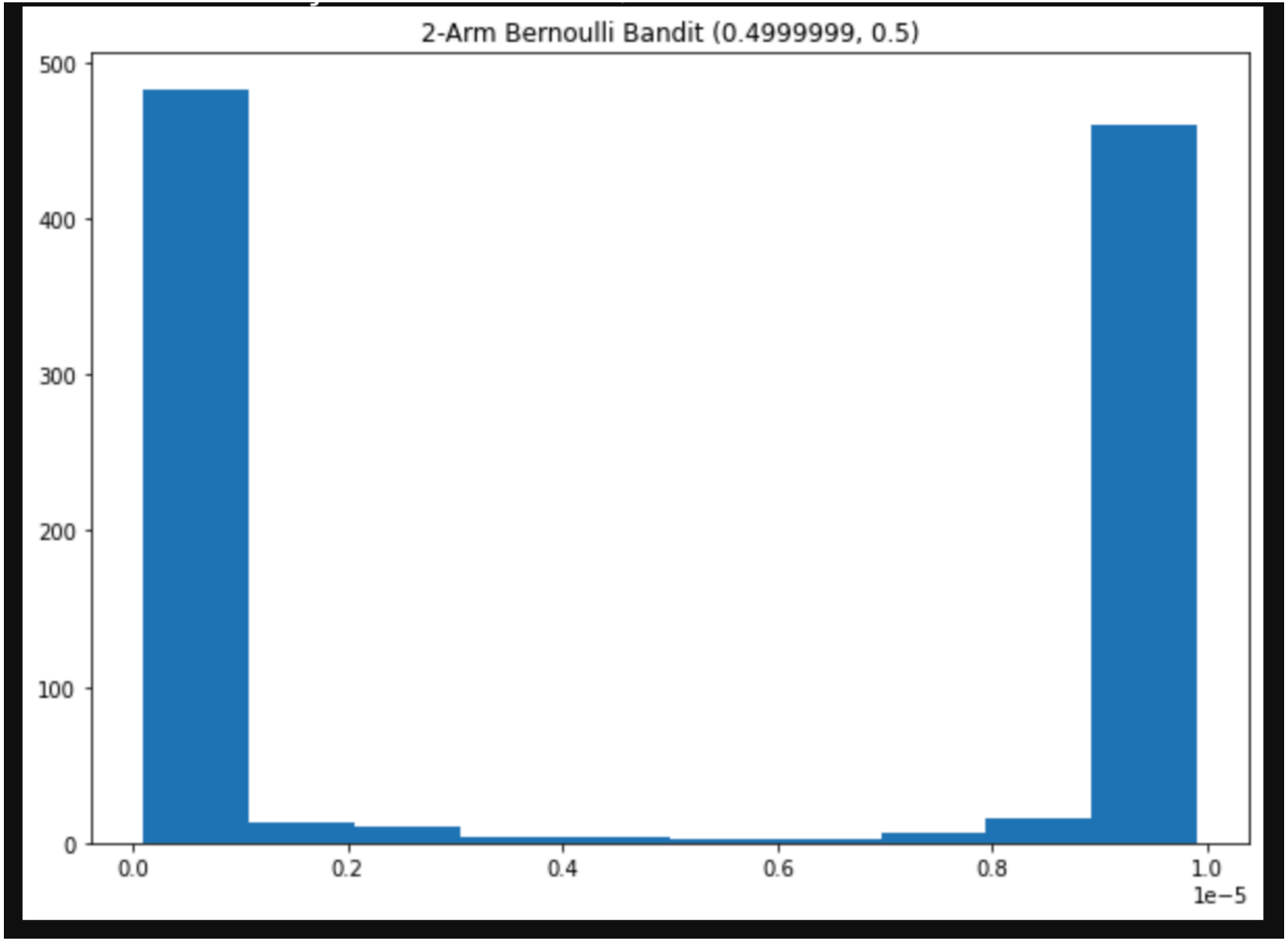
- What does this look like?
  - X-axis is total regret (0-10)
  - Y-axis is # of trials that result in that regret



Simulating this experiment 1000 times

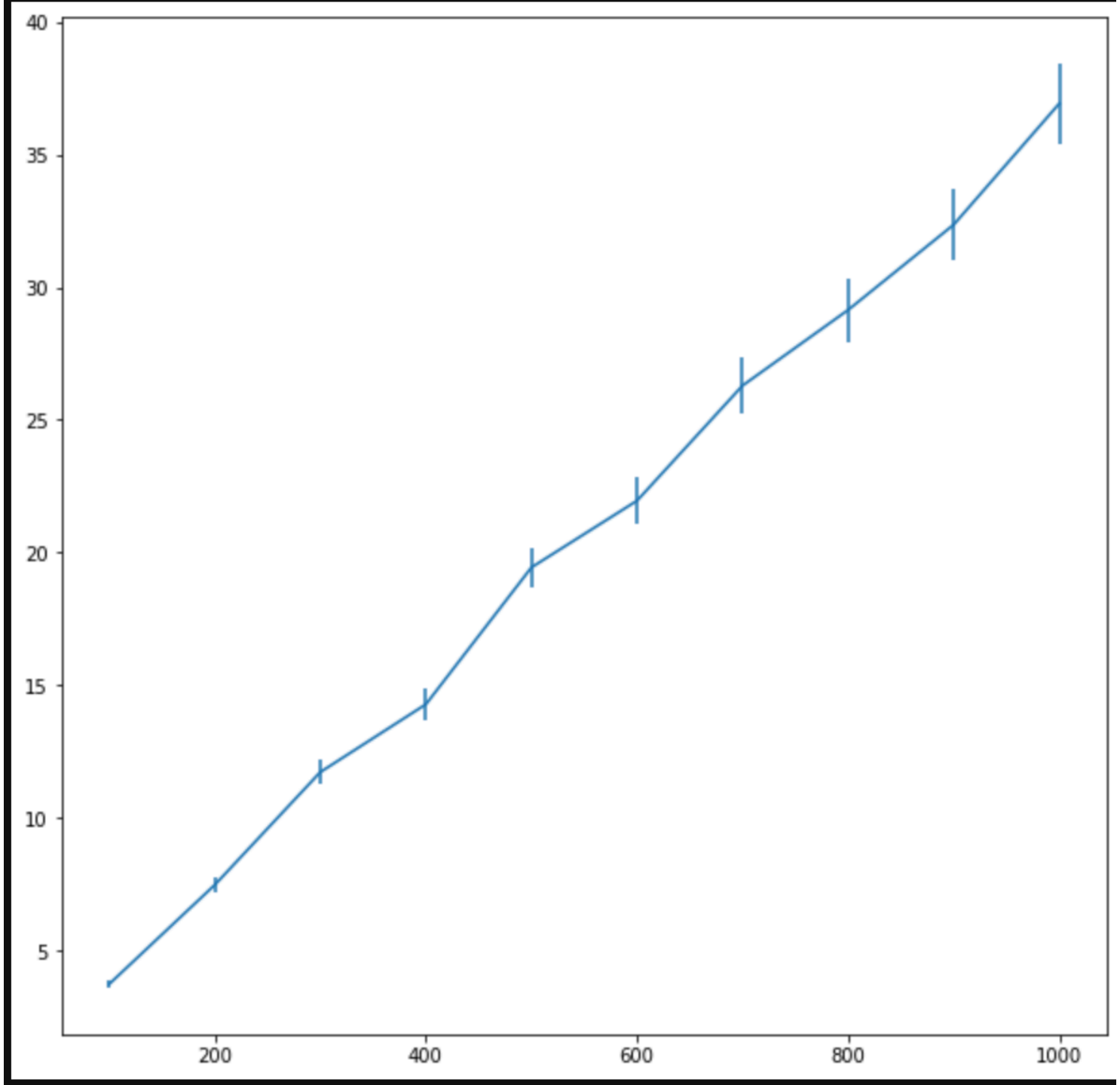



2-Arm Bernoulli Bandit (0.1, 0.9)
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
80


2-Arm Bernoulli Bandit (0.4999999, 0.5)
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
1e−5

How does this scale with horizon?

- A good algorithm is sublinear in n (the horizon)
- More time to learn should result in better performance?
- Sample the Follow-The-Leader for horizon = 100,200,…,1000
  - (1000 trials per)
  - 0.5, 0.6 arms
- What does average regret look like with respect to horizon?


40
35
30
25
20
15
10
5
200
400
600
800
1000