Adversarial Bandits with More Arms than Horizon

Information Theory, Statistics, and Learning Course Project

Amirreza Velae
Amirabbas Afzali

Sharif University of Technology

September 5, 2025

Outline

Real-Life Bandit Algorithm Use Case: Online Ad Placement

- **Context**: E-commerce platforms aim to maximize click-through rate (CTR).
- **Solution**: Use Multi-Armed Bandit (MAB) algorithms like *Thompson Sampling* to balance exploration and exploitation.
- **Impact**: Achieve rapid adaptation and 5–10% CTR uplift within hours.

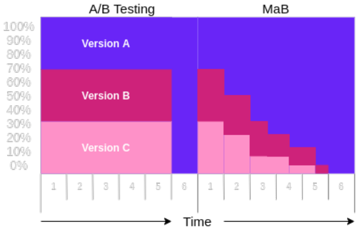


Figure: CTR comparison: MAB vs. A/B test over time

Multi-Armed Bandits

- **Problem:** Sequential decision-making under uncertainty to maximize total reward.
- **Trade-off:**
  - *Explore*: try actions to learn their rewards
  - *Exploit*: choose the best-known action
- **Variants:**
  - *Stochastic*: fixed but unknown reward distributions
  - *Contextual*: use side information (context) per round
- **Use Cases:** Recommendation engines, clinical trials, ad placement

Online Learning vs. Bandits

Online Learning

- *Full feedback*: loss for every action each round
- *Goal*: minimize regret vs. best fixed decision

Bandit Algorithms

- *Partial feedback*: only observe chosen action's reward
- *Goal*: trade off exploration/exploitation to maximize reward

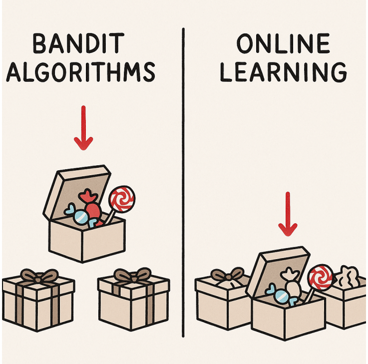


Figure: Comparison of Online Learning and Bandit Algorithms

Exploration vs. Exploitation in Bandit Algorithms

Exploration

- Perturb expert predictions (e.g., randomised sampling)
- Discover under-evaluated strategies and gather data

Exploitation

- Follow-the-Leader: choose the top-performing expert
- Maximise immediate reward based on past performance

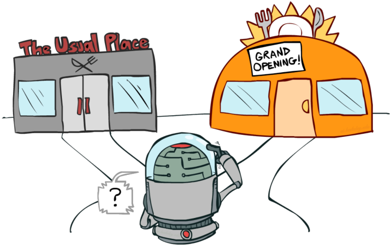


Figure: Trade-off between exploring new experts and exploiting the best expert.

Contextual vs. Canonical Bandits

Canonical (Non-Contextual) Bandits

- K arms, each with unknown reward distribution
- No side information (context) is available
- Example: slot machines with different payout rates

Contextual Bandits

- At each round, a context (feature vector) is observed
- Reward depends on both chosen arm and context
- Example: personalized ads based on user features

Contextual vs. Canonical Bandits

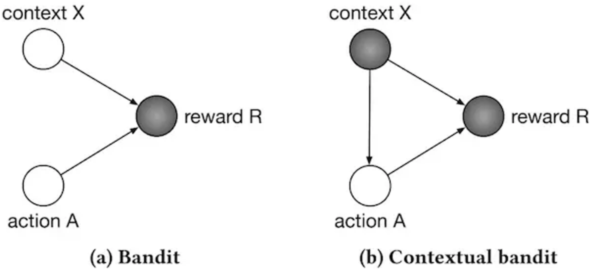


Figure 2: Graphical model notation of the multi-armed bandit and the contextual multi-armed bandit. Nodes indicate random variables, arrows direct conditional dependency, and shaded nodes are observed random variables.

Definition: Regret

Definition

- Regret measures the loss due to not always choosing the best arm.
- For horizon T:

$$R_T = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[r_t]$$

where μ^* is the expected reward of the optimal arm.

Interpretation

- $R_T = 0$: learner always picks the best arm
- Larger regret $\Rightarrow$ poorer learning
- Goal: algorithms with sublinear regret, $R_T = o(T)$

Adversarial Regret

Definition

- Rewards may be chosen by an adversary (non-stationary).
- Regret is defined against the best fixed arm in hindsight:

$$R_T = \max_{i \in [K]} \sum_{t=1}^{T} r_{t,i} - \sum_{t=1}^{T} r_{t,a_t}$$

Interpretation

- Compares learner to the best single arm after T rounds
- Algorithms like EXP3 achieve $R_T = O(\sqrt{TK \log K})$

Generalized Linear Contextual Bandit (GLM-CB)

Context & Arms

- At round $t = 1, \ldots, T$, a context with per-arm features $\{x'_{t,a} \in \mathbb{R}^d : a \in [K]\}$ is observed.
- The learner chooses an arm $a_t \in [K]$ and observes a reward Y_t.

Generalized Linear Model

- Unknown parameter $\theta^* \in \mathbb{R}^d$ and a fixed, strictly increasing link $\mu : \mathbb{R} \to \mathbb{R}$.
- $$\mathbb{E}[\, Y_t \mid x'_{t,a_t} \,] \;=\; \mu\big((x'_{t,a_t})^\top \theta^*\big).$$
- Special cases: linear bandit $\mu(x) = x$; logistic bandit $\mu(x) = \frac{1}{1+e^{-x}}$.

GLM-CB: Optimal Action and Regret

Optimal Action

$$a_t^* = \arg\max_{a \in [K]} \mu\big((x'_{t,a})^\top \theta^*\big).$$

Cumulative Regret of policy π

$$R_T(\pi) := \sum_{t=1}^{T} \Big(\mu\big((x'_{t,a_t^*})^\top \theta^*\big) - \mu\big((x'_{t,a_t})^\top \theta^*\big) \Big).$$

UCB-GLM Algorithm

Algorithm 1 UCB-GLM

Input: the total rounds T, tuning parameter τ and α.
Initialization: randomly choose $a_t \in [K]$ for $t \in [\tau]$, set $V_{\tau+1} = \sum_{i=1}^{\tau} X_t X_t'$
For $t = \tau+1, \tau+2, \ldots, T$ **do**

1. Calculate the maximum-likelihood estimator $\hat{\theta}_t$ by solving the equation
$$\sum_{i=1}^{t-1} (Y_i - \mu(X_i'\theta))X_i = 0 \tag{6}$$
2. Choose $a_t = \mathrm{argmax}_{a \in [K]} \left(X_{t,a}' \hat{\theta}_t + \alpha \left\| X_{t,a} \right\|_{V_t^{-1}} \right)$
3. Observe Y_t, let $X_t \leftarrow X_{t,a_t}$, $V_{t+1} \leftarrow V_t + X_t X_t'$

End For

- For exploitation

UCB-GLM Algorithm [5]

Algorithm 1 UCB-GLM

Input: the total rounds T, tuning parameter τ and α.
Initialization: randomly choose $a_t \in [K]$ for $t \in [\tau]$, set $V_{\tau+1} = \sum_{i=1}^{\tau} X_t X_t'$
For $t = \tau+1, \tau+2, \dots, T$ **do**

1. Calculate the maximum-likelihood estimator $\hat{\theta}_t$ by solving the equation

$$\sum_{i=1}^{t-1} (Y_i - \mu(X_i'\theta))X_i = 0 \tag{6}$$

2. Choose $a_t = \text{argmax}_{a \in [K]} \left(X_{t,a}'\hat{\theta}_t + \alpha \underline{\|X_{t,a}\|_{V_t^{-1}}} \right)$
3. Observe Y_t, let $X_t \leftarrow X_{t,a_t}$, $V_{t+1} \leftarrow V_t + X_t X_t'$

End For

- For exploration

UCB-GLM: Regret Bound

Theorem

Fix $\delta \in (0,1)$. There exists a universal constant $C > 0$ such that running UCB-GLM with

$$\alpha \;=\; \frac{\sigma}{\kappa}\sqrt{\tfrac{d}{2}\log\!\left(1+\tfrac{2T}{d}\right)+\log\tfrac{1}{\delta}}, \qquad \tau \;=\; C\sigma_0^{-2}\left(d+\log\tfrac{1}{\delta}\right)$$

yields, with probability at least $1-2\delta$,

$$R_T \;\le\; \tau \;+\; \frac{2L_\mu \sigma\, d}{\kappa}\;\log\!\left(\tfrac{T}{d\delta}\right)\sqrt{T}.$$

$\rightarrow \tilde{O}(d\sqrt{T})$ regret, independent of the number of arms K.

Our Contributions

1. **Problem Setting**
   - A finite many-armed bandit problem.
   - A history-dependent adversarial strategy.
2. **Bounds & Results**
   - A sub-linear regret bound in the online learning setting.
   - A lower bound in the general Canonical High-Arm Regime ($K \geq T$) (Ignoring the structure of the problem).
   - An Algorithm with a sub-linear regret bound in the adversarial setting.

Problem Setting

We consider a *finite many-armed* bandit problem:

- **Action Space:** A set of K arms, where $K \gg T$.
- **Reward Structure:** The reward for each arm is bounded in $[0, 1]$.
- **Learner's Objective:** Select a sequence of arms $\{a_t\}_{t=1}^{T}$ to maximize the cumulative reward.

Adversary's Strategy: Uniform Unseen Arm Rewards

Adversary's behavior in our setting:

- The adversary observes the learner's history $\{a_1, \ldots, a_{t-1}\}$ and acts adaptively.
- All **unseen arms** are grouped into a single abstract arm.
- The adversary then assigns a common reward $r_t^{\text{unseen}} \in [0, 1]$ to all unseen arms uniformly.
- This strategy dynamically shapes the reward structure:

$$r_t(a) = \begin{cases} r_t^{\text{seen}}(a) & \text{if } a \in \{a_1, \ldots, a_{t-1}\} \\ r_t^{\text{unseen}} & \text{otherwise} \end{cases}$$

Regret Definition

- At each round $t \in [T]$, the player observes h_{t-1} and chooses a policy $P_t \in \mathbb{R}^{k_{t+1}}$.
- Then, the player selects action $a_t = i$ with probability $P_{t,i}$, for all $i \in [k_{t+1}]$.

Again, Definition of Adversarial Regret:

$$R_T(\pi, x) = \max_{i \in [k_{T+1}]} \sum_{t=1}^{T} x_{t,i} - \mathbb{E}\left[\sum_{t=1}^{T} x_{t,a_t}\right]$$

Goal & roadmap

Claim (minimax lower bound).

$$\inf_{\text{alg}} \sup_{\ell_{1:T}\in[0,1]^{K\times T}} \mathbb{E}[\mathrm{Reg}_T] \;\ge\; c\,\min\{\sqrt{KT},\; T\}$$

for a universal constant $c > 0$.

Proof roadmap.

1. **Yao's principle:** analyze any *deterministic* learner under a chosen input distribution [6].
2. **Hard environment:** one hidden "good" arm with advantage ε (Bernoulli rewards).
3. **Testing step:** distinguish *null* vs. "good arm i" via history; control via TV/Pinsker or Bretagnolle–Huber, and compute KL with the bandit KL chain rule.
4. **Tune ε:** set $\varepsilon \asymp \sqrt{K/T}$ to get $\mathbb{E}[\mathrm{Reg}_T] \gtrsim \sqrt{KT}$.

Yao's minimax & Testing–TV

Theorem (Yao [6])

For any loss functional $L(\mathsf{Alg}, x)$*,*

$$\max_{\mu} \min_{\textit{det } \mathsf{A}} \mathbb{E}_{x\sim\mu}[L(\mathsf{A}, x)] = \min_{\textit{rand } \mathsf{Alg}} \max_{x} \mathbb{E}[L(\mathsf{Alg}, x)].$$

Theorem (Optimal testing $\leftrightarrow$ total variation)

For distributions P, Q *with densities* p, q*,*

$$\alpha^\star(P, Q) = \min_{\phi}\{P[\phi = 1] + Q[\phi = 0]\} = 1 - \mathrm{TV}(P, Q),$$

The optimal test is the likelihood-ratio rule $\phi^\star = \mathbf{1}\{p \geq q\}$ *(Neyman–Pearson).*

Pinsker & Bretagnolle–Huber (at a glance)

Theorem

Pinsker [3]

$$\mathrm{TV}(P, Q) \;\le\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}(P\|Q)}.$$

Use: bound differences of expectations of bounded stats via TV.

Bernoulli KL plug-in

For $p = \frac{1}{2}$, $q = \frac{1}{2} + \varepsilon$,

$$d\Big(\tfrac{1}{2}\Big\|\tfrac{1}{2} + \varepsilon\Big) = \tfrac{1}{2}\log\frac{1}{1-4\varepsilon^2} = 2\varepsilon^2 + O(\varepsilon^4).$$

For $\varepsilon \le \frac{1}{4}$,

$$-\log(1-4\varepsilon^2) \;\le\; 16\ln\tfrac{4}{3}\;\varepsilon^2 \quad\Rightarrow\quad d\Big(\tfrac{1}{2}\Big\|\tfrac{1}{2}+\varepsilon\Big) \;\le\; 8\ln\tfrac{4}{3}\;\varepsilon^2.$$

TV $\Rightarrow$ expectations; KL chain rule

Lemma (TV bound)

If $f : H_T \to [0, M]$, *then* $\quad |\mathbb{E}_P f - \mathbb{E}_Q f| \le M \,\mathrm{TV}(P, Q)$.

Applied to $f = T_i(T) \in [0, T]$:

$$\mathbb{E}_{P_i} T_i(T) \;\le\; \mathbb{E}_{P_0} T_i(T) \;+\; T\,\mathrm{TV}(P_i, P_0).$$

Lemma (KL decomposition [4])

If P, Q *differ only on arm* i, *then*

$$\mathrm{KL}(P \| Q) \;=\; \mathbb{E}_Q[T_i(T)]\; D(P_i \| Q_i).$$

Under the symmetric null Q: $\;\mathbb{E}_Q[T_i(T)] = T/K$.

Environment. Sample a hidden good arm $I \sim \mathrm{Unif}\{1, \ldots, K\}$. For each round t and arm j,

$$X_{tj} \sim \begin{cases} \mathrm{Bernoulli}(\frac{1}{2} + \varepsilon), & j = I, \\ \mathrm{Bernoulli}(\frac{1}{2}), & j \neq I, \end{cases} \quad \text{independently.}$$

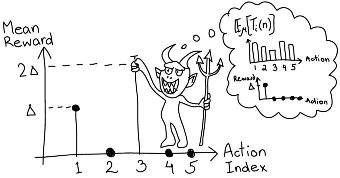
Let P_i be the law given $I = i$, and P_0 the *null* (all arms Bernoulli($\frac{1}{2}$)).

Regret identity.

$$\mathbb{E}_{P_i}[\mathrm{Reg}_T] \;\geq\; \varepsilon\big(T - \mathbb{E}_{P_i}[T_i(T)]\big).$$

Symmetry (under P_0).

$$\mathbb{E}_{P_0}[T_i(T)] = T/K \quad \text{for all } i.$$



One hidden "good" arm with gap ε; missing it drives regret.

TV $\Rightarrow$ Pinsker $\Rightarrow$ KL

TV bound For $T_i(T) \in [0, T]$,

$$\mathbb{E}_{P_i}[T_i(T)] \;\le\; \mathbb{E}_{P_0}[T_i(T)] \;+\; T\,\mathrm{TV}(P_i, P_0).$$

Pinsker + chain rule + symmetry

$$\begin{aligned}\mathrm{TV}(P_i, P_0) \;\le\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}(P_0 \| P_i)} &= \;\sqrt{\tfrac{1}{2}\,\mathbb{E}_{P_0}[T_i(T)]\; d(\tfrac{1}{2} \| \tfrac{1}{2} + \varepsilon)} \\ &= \;\sqrt{\frac{1}{2}\,\frac{T}{K}\; d(\tfrac{1}{2} \| \tfrac{1}{2} + \varepsilon)}.\end{aligned}$$

Conclusion.

$$\boxed{\mathbb{E}_{P_i}[T_i(T)] \;\le\; \frac{T}{K} \;+\; \frac{T}{2}\sqrt{\frac{T}{K} \cdot \big(-\log(1 - 4\varepsilon^2)\big)}}.$$

Putting it together (constants-ready)

Using the regret identity,

$$\mathbb{E}_{P_i}[\mathrm{Reg}_T] \;\geq\; \varepsilon \left(T - \frac{T}{K} - \frac{T}{2}\sqrt{\frac{T}{K}\left(-\log(1-4\varepsilon^2)\right)} \right).$$

For $\varepsilon \leq \frac{1}{4}$ and $-\log(1-4\varepsilon^2) \leq 8\varepsilon^2$,

$$\mathbb{E}_{P_i}[\mathrm{Reg}_T] \;\geq\; \varepsilon T\Big(1 - \tfrac{1}{K}\Big) \;-\; 2\,T\varepsilon^2\sqrt{\tfrac{T}{K}}.$$

Pick $\varepsilon = \dfrac{1}{4}\min\left\{1, \sqrt{\tfrac{K}{T}}\right\} \Rightarrow \mathbb{E}[\mathrm{Reg}_T] \;\geq\; c\min\{\sqrt{KT},\; T\}.$

Canonical high-arm regime ($K \geq T$)

Regime. More arms than time ($T \leq K$).

Minimax picture (oblivious).

$$\inf_{\text{alg}} \sup_{\ell_{1:T} \in [0,1]^{K \times T}} \mathbb{E}[\mathrm{Reg}_T] \;\geq\; c \min\{\sqrt{KT},\, T\}$$

$$\Rightarrow \quad K \geq T: \;\; \mathbb{E}[\mathrm{Reg}_T] \;\geq\; c\, T.$$

(Bartlett gives $c = 1/18$.) [1]

Intuition. With T rounds and K arms, most arms are unseen; a single better arm is rarely found ($\mathbb{E}_{P_0}[T_i(T)] = T/K \ll 1$), so regret is the price of missing it.

Upper bounds & takeaways

Upper bound EXP3.

$$\mathbb{E}[\mathrm{Reg}_T] \;\le\; \min\left\{\, T,\; C\sqrt{KT\log K}\,\right\} \qquad [2].$$

When $K \ge T$, the trivial cap T dominates:

$$\boxed{\mathbb{E}[\mathrm{Reg}_T] \;=\; \Theta(T) \;\; \text{for } K \ge T.}$$

Practical notes.

- No algorithm beats $\Omega(T)$ worst-case without extra structure.

- To improve: inject structure (contexts/experts), prune arms, or accept linear regret guarantees.

History-dependent bandit

Pool unseen arms. At round t:

$$S_{t-1} = \{a_1, \ldots, a_{t-1}\}, \quad \mathcal{A}_t = S_{t-1} \cup \{U\}, \quad |\mathcal{A}_t| \le t.$$

Rewards:

$$r_t(a) = \begin{cases} r_t^{\text{seen}}(a), & a \in S_{t-1}, \\ r_t^{\text{unseen}}, & a = U. \end{cases}$$

If U is played: spawn a^{new}, reveal r_t^{unseen}, set $S_t = S_{t-1} \cup \{a^{\text{new}}\}$.

Learner. Run EXP3 on $\mathcal{A}_t$; on spawn initialize

$$w_t(a^{\text{new}}) = \frac{1}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} w_t(a).$$

A History-Dependent Bandit Interpretation

This setting can be seen as a **nonstationary** bandit problem:

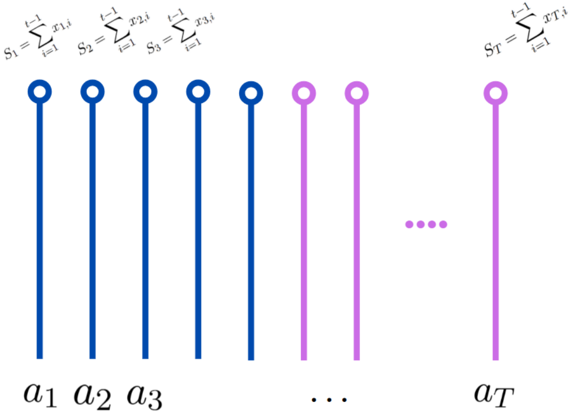
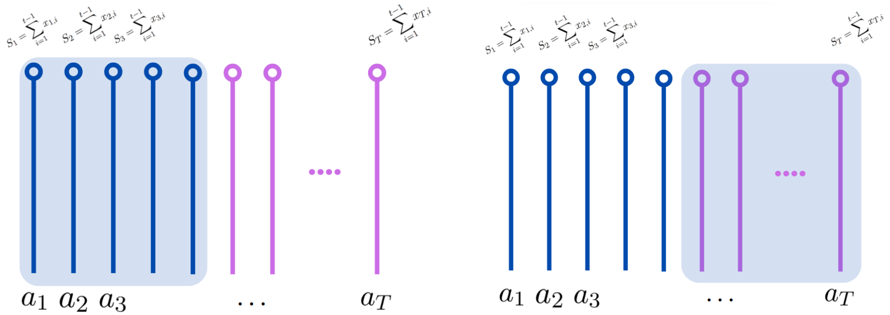


Figure: Multi-Armed Bandit

A History-Dependent Bandit Interpretation

This setting can be seen as a **nonstationary** bandit problem:



(a) Option 1   (b) Option 2

Figure: Illustration of two existing action options at time step t, representing dynamic arm sets in a history-dependent bandit setting.

Reduction Lemma

Lemma. For every original arm j there exists $b_j \in S_T \cup \{U\}$ such that

$$\sum_{t=1}^{T} r_t(j) \;=\; \sum_{t=1}^{T} \tilde{r}_t(b_j), \quad \tilde{r}_t(b) = \begin{cases} r_t^{\text{unseen}}, & t < \tau_b, \\ r_t^{\text{seen}}(b), & t \geq \tau_b, \end{cases}$$

where τ_b is the first time b appears. Hence

$$\max_{j \in [K]} \sum_{t=1}^{T} r_t(j) \;=\; \max_{b \in S_T \cup \{U\}} \sum_{t=1}^{T} \tilde{r}_t(b), \qquad |S_T \cup \{U\}| \leq T{+}1.$$

Intuition: Any original arm is "unseen" until its first pull, so its path equals "play U until it appears, then play it forever."

Consequence for Regret

Apply EXP3 on the reduced set.

- *Prior:* average-weight init $\Rightarrow$ penalty $\log|\mathcal{A}_{\tau_b}| \leq \log(T+1)$.
- *Variance:* use implicit (or explicit) exploration to control the $1/p$ terms.

Result. Against a non-anticipating adaptive adversary with rewards in $[0, 1]$,

$$\mathbb{E}[\mathrm{Regret}_T] = O\Big(\sqrt{T \log(T+1)}\Big).$$

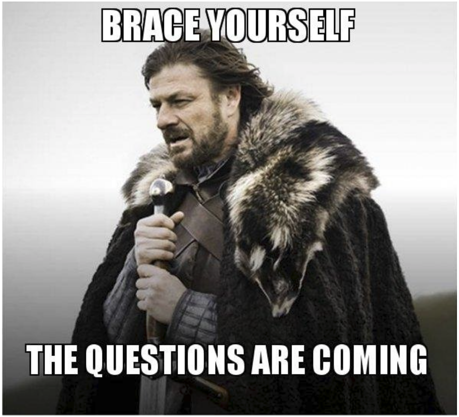
Takeaway. The "pool unseen as one arm" reduction shrinks the effective comparator set to $\leq T+1$, yielding the usual $\tilde{O}(\sqrt{T})$ regret with off-the-shelf adversarial bandit algorithms.

Acknowledgment

- Special thanks to Mr. **Sarzaeem** for his valuable suggestion regarding the setting.
- Thanks to Mr. **Zinati** for advice on this presentation.

Q & A

Questions?
Feel free to ask for clarifications, share feedback, or start a discussion.



References I

[1] Peter Bartlett. Adversarial bandits (lecture notes). Course notes, Stat 260/CS 294-102, UC Berkeley, 2014. URL `https://www.stat.berkeley.edu/~bartlett/courses/2014fall-cs294stat260/lectures/bandit-adversarial-notes.pdf`.

[2] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024. URL `https://www.nowpublishers.com/article/Details/MAL-024`.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006. ISBN 9780471241959. doi: 10.1002/047174882X. URL `https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X`.

References II

[4] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401. URL `https://tor-lattimore.com/downloads/book/book.pdf`.

[5] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits, 2017. URL `https://arxiv.org/abs/1703.00048`.

[6] Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.