Adversarial Bandits with More Arms than Horizon

Information Theory, Statistics, and Learning

Amirreza Velae* Amirabbas Afzali†

September 5, 2025

Abstract

We study adversarial multi–armed bandits in the *high–arm* regime in which the number of actions K can be comparable to or exceed the horizon T. For the canonical setting with bandit feedback and losses in $[0, 1]$, we provide a concise information–theoretic proof of the minimax lower bound

$$\Omega\big(\min\{\sqrt{KT},\, T\}\big)$$

on the expected regret [2, 4, 14]. The argument is phrased as a binary hypothesis test between a null model (all arms $\mathrm{Ber}(1/2)$) and a single "good–arm" alternative ($\mathrm{Ber}(1/2+\varepsilon)$ on one arm), and combines the testing–total variation identity [20] with Pinsker's and Bretagnolle–Huber inequalities [3, 8] and a Kullback–Leibler (KL) divergence *chain rule* tailored to bandit feedback [2, 4, 14]. This matches, up to logarithmic factors, the classical EXP3 upper bounds [2, 4]. We then exploit structure in a history–dependent model where the adversary assigns a common reward to all yet–unseen arms. Pooling unseen arms into one abstract action reduces the effective comparator set to at most $T+1$ items, which yields an EXP3–style procedure with regret $O\big(\sqrt{T\log(T+1)}\big)$ under bandit feedback (cf. analyses for dynamic/sleeping action sets [11]). The rates follow by balancing KL $\sim (T/K)\varepsilon^2$ against regret $\sim \varepsilon T$, giving the inevitable choice $\varepsilon \asymp \sqrt{K/T}$ [14].

1 Introduction

Motivation. Sequential decision problems with partial feedback (bandits) arise in high–throughput systems such as ad selection, recommendation, and online routing, where exploration must be traded against exploitation and the environment can be nonstationary or adversarial [4, 14]. In such applications the action set is often large ($K \gg T$), so guarantees must scale correctly in both K and T and avoid stochastic assumptions when they are unwarranted.

In the adversarial K–armed bandit, tight rates are known up to logarithmic factors: the EXP3 family achieves $O(\sqrt{KT\log K})$, and the minimax lower bound is $\Omega(\min\{\sqrt{KT}, T\})$; in particular, regret is linear in the extreme high–arm regime $K \geq T$ [2, 4, 14]. These results delineate what is achievable without additional structure.

Contributions.

- **Model and scope.** We formalize the high–arm adversarial bandit with bandit feedback and an oblivious (loss–sequence) adversary for the lower bound, and a history–dependent but non–anticipating adversary for the structured result in which unseen arms are symmetric (related in spirit to sleeping/dynamic action models [11]).
- **Lower bound via hypothesis testing.** We give a compact proof of the canonical minimax lower bound $\Omega\big(\min\{\sqrt{KT}, T\}\big)$ by reducing regret to a binary test between a null and "one good arm" alternative. The proof uses the testing–TV equality [20], Pinsker's and Bretagnolle–Huber inequalities [3, 8] to control testing error, and a bandit KL chain rule to evaluate the divergence [2, 4, 14].
- **Structured algorithm under pooled unseen arms.** When the adversary treats all unseen arms identically, pooling them into a single abstract arm shrinks the effective comparator set to $\leq T+1$. Running an EXP3–style learner on the evolving action set—with average–weight initialization for newly spawned arms—yields $O(\sqrt{T\log(T+1)})$ regret under bandit feedback [cf. 2, 11].

2 Background

2.1 Multi-armed bandits

The (stochastic or adversarial) multi-armed bandit (MAB) formalizes sequential decision making with partial feedback: at each round t, a learner selects an arm $a_t \in [K]$ and only observes the loss (or reward) of that arm. The

*Department of Electrical Engineering, Sharif University of Technology. E-mail: `amirreza.velae@ee.sharif.edu`

†Department of Electrical Engineering, Sharif University of Technology. E-mail: `afzali@ee.sharif.ac.ir`

model dates to early work on sequential experimental design [18] and has since become a central abstraction for online optimization under uncertainty. In the stochastic case, each arm i yields i.i.d. rewards from an unknown distribution; in the adversarial case, the loss vectors $(\ell_t(i))_{i=1}^K \in [0,1]^K$ may be chosen by an oblivious or adaptive adversary, with the learner observing only $\ell_t(a_t)$. Classical results establish near-matching minimax rates: EXP3 attains $O(\sqrt{KT \log K})$ expected regret in the adversarial setting, and $\Omega(\sqrt{KT})$ is unavoidable (up to constants/logs). [2, 4, 14] :contentReference[oaicite:0]index=0

2.2 Online learning vs. bandits

Online learning (experts/online convex optimization) differs from bandits in the feedback model: full-information learners observe the entire loss vector $\ell_t(\cdot)$ each round, whereas bandit learners only see $\ell_t(a_t)$. Consequently, the best-possible regret scales as $O(\sqrt{T \log K})$ with full feedback but as $O(\sqrt{KT})$ with bandit feedback—precisely due to the information bottleneck. Standard references formalize these regimes and regret notions (external/pseudo-regret), as well as reductions connecting experts and bandits. [4, 6, 19] :contentReference[oaicite:1]index=1

2.3 Contextual vs. canonical bandits

In *canonical* (non-contextual) bandits, arm losses depend only on the arm and time. *Contextual* bandits augment each round with side information x_t and allow the loss of arm i to depend on (x_t, i), so the learner competes with a policy class mapping contexts to arms. Foundational algorithms include EXP4 (policy-based), Epoch-Greedy (supervised-oracle-based), and efficient linear models such as LinUCB; subsequent work gave statistically optimal, oracle-efficient approaches. [1, 4, 7, 13, 15] :contentReference[oaicite:2]index=2

3 Problem Setup

3.1 High-arm regime and notation

We study a K-armed bandit with horizon T. At round $t = 1, \dots, T$, the learner selects $a_t \in [K]$ and incurs loss $\ell_t(a_t) \in [0,1]$. Let A_t denote the action random variable, and let $H_t = (A_1, \ell_1(A_1), \dots, A_t, \ell_t(A_t))$ be the history. We emphasize the *high-arm* regime $K \gg T$ (often $K \geq T$), common in cold-start applications where the catalog of actions is large relative to the time budget. Unless stated otherwise, the adversary is *oblivious* for the lower-bound analysis (fixing $(\ell_t(i))_{t,i}$ in advance) and *non-anticipating* in our structured model (losses may depend on past but not on current randomization). The feedback is bandit: only $\ell_t(a_t)$ is observed. [2, 4, 14] :contentReference[oaicite:3]index=3

3.2 Regret definitions

For a (possibly randomized) learner π, the *adversarial (external) regret* against the best static arm in hindsight is

$$\mathrm{Reg}_T(\pi; \ell_{1:T}) = \sum_{t=1}^{T} \ell_t(A_t) \; - \; \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i).$$

We will primarily analyze $\mathbb{E}[\mathrm{Reg}_T]$ where the expectation is over the learner's randomness (and any randomness in the environment when we invoke Yao's minimax principle). For context, in the stochastic MAB with arm means (μ_i), the standard regret is $R_T = T\mu^\star - \sum_{t=1}^{T} \mathbb{E}[r_t(A_t)]$, with $\mu^\star = \max_i \mu_i$ and $r_t = 1 - \ell_t$. [4, 6, 14, 18] :contentReference[oaicite:4]index=4

4 Adversary Model (History-Dependent)

We consider a *non-anticipating, history-dependent* adversary. At time t, the loss vector $\ell_t \in [0,1]^K$ may depend on the prior interaction history H_{t-1}, but not on the learner's randomized action, which is revealed only after the adversary commits to ℓ_t [14]. This model properly captures dynamic, nonstationary environments while ensuring that regret remains well-defined.

4.1 Pooling unseen arms

Under the assumption that the adversary treats all *unseen* arms identically, we can introduce an abstract arm U representing the entire unseen set. Formally, let

$$S_{t-1} = \{a_1, \dots, a_{t-1}\} \quad \text{and} \quad \mathcal{A}_t = S_{t-1} \cup \{U\}.$$

For each t, losses are given by:

$$\ell_t(i) = \begin{cases} \ell_t^{\text{seen}}(i; H_{t-1}), & i \in S_{t-1}, \\ \ell_t^{\text{unseen}}(H_{t-1}), & i \notin S_{t-1}. \end{cases}$$

If the learner selects U, a fresh unseen arm is spawned, its identity revealed along with its loss, and it enters S_t. This reduction preserves loss sequences and guarantees that the effective action set size remains bounded by $T + 1$, enabling direct application of adversarial bandit algorithms (like EXP3) with standard regret analysis [14].

5 Adversary Model (History-Dependent)

We consider a *non-anticipating, history-dependent* adversary. At time t, the loss vector $\ell_t \in [0,1]^K$ may depend on the prior interaction history H_{t-1}, but not on the learner's randomized action, which is revealed only after the adversary commits to ℓ_t [14]. This model properly captures dynamic, nonstationary environments while ensuring that regret remains well-defined.

5.1 Pooling unseen arms

Under the assumption that the adversary treats all *unseen* arms identically, we can introduce an abstract arm U representing the entire unseen set. Formally, let

$$S_{t-1} = \{a_1, \dots, a_{t-1}\} \quad \text{and} \quad \mathcal{A}_t = S_{t-1} \cup \{U\}.$$

For each t, losses are given by:

$$\ell_t(i) = \begin{cases} \ell_t^{\text{seen}}(i; H_{t-1}), & i \in S_{t-1}, \\ \ell_t^{\text{unseen}}(H_{t-1}), & i \notin S_{t-1}. \end{cases}$$

If the learner selects U, a fresh unseen arm is spawned, its identity revealed along with its loss, and it enters S_t. This reduction preserves loss sequences and guarantees that the effective action set size remains bounded by $T+1$, enabling direct application of adversarial bandit algorithms (like EXP3) with standard regret analysis [14].

6 Information-Theoretic Tools

6.1 Yao's minimax principle

Yao's principle transforms a minimax lower bound for randomized algorithms into a lower bound for deterministic algorithms under a randomized input (hard distribution). Formally,

$$\inf_{\text{randomized alg}} \sup_x \mathbb{E}[L] \;\geq\; \sup_\mu \inf_{\text{deterministic alg}} \mathbb{E}_{x\sim\mu}[L].$$

We apply this by selecting a distribution over loss sequences that is hard for any deterministic algorithm [14, Chapter 14].

6.2 Testing and total variation

In binary hypothesis testing between distributions P and Q, the optimal error probability equals $1 - \mathrm{TV}(P,Q)$, where TV is total variation distance. Importantly, for any $f : \mathcal{H} \to [0, M]$,

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq M\,\mathrm{TV}(P,Q).$$

We use this to bound differences in expectations of bounded statistics (e.g., arm pull counts) across two bandit environments.

6.3 Pinsker and Bretagnolle–Huber inequalities

Pinsker's inequality provides:

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

which is tight for small divergences [5, 9]. We employ Pinsker to control expectation differences.

6.4 Bandit KL chain rule

When two environments differ only on arm i, a fixed deterministic policy yields:

$$\mathrm{KL}(P\|Q) = \mathbb{E}_Q[N_i] \cdot d(p_i\|q_i),$$

where N_i is the number of times arm i is pulled under Q, and d is the single-round divergence (e.g., Bernoulli KL). This identity decomposes the divergence along the interaction and is crucial in relating KL to expected pulls [14].

7 Minimax Lower Bound

Theorem 1 (Canonical adversarial lower bound)**.** *For all $K \geq 2$, $T \geq 1$, any bandit algorithm (possibly randomized) suffers*

$$\mathbb{E}[\mathrm{Reg}_T] \;\geq\; c\,\min\{\sqrt{KT},\, T\}$$

for a universal constant $c > 0$.

Proof overview. We argue by Yao's minimax principle: it suffices to exhibit a distribution over loss sequences for which every deterministic learner suffers the stated expected regret. We use a binary hypothesis testing reduction between a *null* environment and a *single–good–arm* alternative, and control distinguishability via total variation and KL divergence (Pinsker or Bretagnolle–Huber). The bandit KL *chain rule* translates indistinguishability into an upper bound on the expected number of pulls of the good arm, which, via a simple one–line regret identity, yields the lower bound after tuning a gap parameter ε. ☐

7.1 Hard environment

Fix a gap parameter $\varepsilon \in (0, \frac{1}{4}]$. Sample a hidden index $I \sim \mathrm{Unif}\{1, \dots, K\}$. At each round $t \in [T]$ and for each arm $j \in [K]$, generate an i.i.d. Bernoulli loss

$$X_{t,j} \;\sim\; \begin{cases} \mathrm{Ber}(\frac{1}{2} - \varepsilon), & j = I, \\ \mathrm{Ber}(\frac{1}{2}), & j \neq I, \end{cases}$$

and let the learner observe only X_{t,a_t} for its chosen arm a_t. Denote by P_i the law of the full history H_T conditional on $I = i$, and by P_0 the *null* law under which all arms are $\mathrm{Ber}(\frac{1}{2})$ (so I is irrelevant). This construction is standard in adversarial bandit lower bounds. [14, Ch. 15] [2].

7.2 Regret identity

Let $N_i = \sum_{t=1}^{T} \mathbf{1}\{a_t = i\}$ be the (random) number of pulls of arm i. Under P_i, the best fixed arm in hindsight is i with expected cumulative loss $(\frac{1}{2}-\varepsilon)T$, while the learner's expected cumulative loss is $\frac{1}{2}T - \varepsilon\, \mathbb{E}_{P_i}[N_i]$. Therefore,

$$\mathbb{E}_{P_i}[\mathrm{Reg}_T] \;\geq\; \varepsilon\Big(T - \mathbb{E}_{P_i}[N_i]\Big). \tag{1}$$

This is the fundamental "price of not identifying the good arm" inequality. [14, Sec. 15.2].

7.3 Route A: Pinsker (expectations)

Step A1 (TV controls bounded statistics). For any $f : \mathcal{H}_T \to [0, M]$ and distributions P, Q on histories,

$$\big|\mathbb{E}_P f - \mathbb{E}_Q f\big| \;\leq\; M\, \mathrm{TV}(P, Q).$$

Apply to $f = N_i \in [0, T]$ and $(P, Q) = (P_i, P_0)$:

$$\mathbb{E}_{P_i}[N_i] \;\leq\; \mathbb{E}_{P_0}[N_i] \;+\; T\, \mathrm{TV}(P_i, P_0). \tag{2}$$

The identity $\mathbb{E}_{P_0}[N_i] = T/K$ holds by symmetry under P_0.

Step A2 (Pinsker + bandit KL chain rule). Pinsker's inequality gives $\mathrm{TV}(P_i, P_0) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P_0\|P_i)}$. By the *bandit KL chain rule*, when environments differ only on arm i,

$$\mathrm{KL}(P_0\|P_i) \;=\; \mathbb{E}_{P_0}[N_i]\cdot d\big(\tfrac{1}{2} \,\big\|\, \tfrac{1}{2} - \varepsilon\big) \;=\; \frac{T}{K}\cdot d\big(\tfrac{1}{2} \,\big\|\, \tfrac{1}{2} - \varepsilon\big)\,,$$

where $d(\cdot\|\cdot)$ is the one–step (Bernoulli) KL divergence. For Bernoulli parameters $p, q \in (0, 1)$, $d(p\|q) = p \log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$, and in particular

$$d\big(\tfrac{1}{2} \,\big\|\, \tfrac{1}{2} - \varepsilon\big) \;=\; \tfrac{1}{2}\log\Big(\tfrac{1}{1-4\varepsilon^2}\Big).$$

Combining with (2),

$$\mathbb{E}_{P_i}[N_i] \;\leq\; \frac{T}{K} \;+\; \frac{T}{2}\sqrt{\frac{T}{K}\cdot\big(-\log(1-4\varepsilon^2)\big)}. \tag{3}$$

Now use $-\log(1-4\varepsilon^2) \leq 8\varepsilon^2$ for $\varepsilon \leq \frac{1}{4}$ to obtain

$$\mathbb{E}_{P_i}[N_i] \;\leq\; \frac{T}{K} \;+\; 2T\varepsilon\sqrt{\frac{T}{K}}.$$

Insert this into (1):

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

Finally, choose $\varepsilon = \min\{\frac{1}{4},\ c_0\sqrt{K/T}\}$ with a small numerical c_0 to balance the two terms (e.g., $c_0 = \frac{1}{4}$), yielding

$$\mathbb{E}_{P_i}[\mathrm{Reg}_T] \;\gtrsim\; \min\{\sqrt{KT},\, T\},$$

and hence the minimax lower bound by Yao's principle. [8, 14, Ch. 2 (Pinsker); Ch. 15 (KL chain rule, lower bound)].

7.4 Constants and discussion

The constant c can be traced through the inequalities above; classical treatments (and refined analyses) report absolute constants of this form and show tightness (up to logs) against EXP3's $O(\sqrt{KT\log K})$ upper bound and against the trivial cap T. In the *high-arm* regime $K \geq T$, the bound simplifies to $\mathbb{E}[\mathrm{Reg}_T] \geq cT$, i.e., linear regret is information-theoretically unavoidable without additional structure. See Auer et al. [2] for the original nonstochastic formulation and Bubeck and Cesa-Bianchi [4], Lattimore and Szepesvári [14] for modern expositions; see also Gerchinovitz and Lattimore [10] for refined lower bounds matching several sharpened upper bounds (e.g., high-probability or variation-dependent forms).

8 Algorithm in the Structured Setting

8.1 Reduction lemma (pooling unseen ⇒ at most $T{+}1$ comparators)

Let $S_{t-1} = \{a_1, \ldots, a_{t-1}\}$ be the set of *distinct* arms pulled before t, and assume the history-dependent, non-anticipating adversary assigns a common loss $\ell_t^{\text{unseen}}(H_{t-1})$ to all arms not in S_{t-1}, while seen arms $i \in S_{t-1}$ receive $\ell_t^{\text{seen}}(i; H_{t-1})$ (Sec. 5). Introduce a single abstract arm U representing the entire set of unseen arms, and define

$$\mathcal{A}_t = S_{t-1} \cup \{U\}, \qquad |\mathcal{A}_t| \leq t.$$

Lemma 2 (Reduction)**.** *For every original arm $j \in [K]$ there exists $b_j \in S_T \cup \{U\}$ such that*

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

where τ_b is the (random) first time b appears in S_t (and $\tau_U = \infty$). Consequently,

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

Proof sketch. Fix an original arm j and let τ_j be its reveal time (the first round it is pulled; $\tau_j = \infty$ if never pulled). Before τ_j, j is indistinguishable from any unseen arm by assumption, hence $\ell_t(j) = \ell_t^{\text{unseen}}$ for $t < \tau_j$. At time τ_j, the abstract arm U *spawns* the concrete arm b_j that coincides with j thereafter; thus for $t \geq \tau_j$, $\ell_t(j) = \ell_t^{\text{seen}}(b_j)$. If j is never pulled, take $b_j = U$. Summing over t gives the pathwise identity and therefore the equality of benchmarks. ☐

8.2 EXP3 on evolving action sets

We run an exponential-weights bandit algorithm (EXP3/EXP3-IX) on the evolving set $\mathcal{A}_t$ [2, 4]. The only nonstandard ingredient is the *average-weight initialization* for newly spawned arms, which preserves the potential $\log \sum_{a \in \mathcal{A}_t} w_t(a)$ so that the prior-cost term scales with $\log |\mathcal{A}_t| \leq \log(T+1)$, exactly as in fixed-$K$ analyses of Hedge/EXP3 [4]. For variance control we recommend the implicit-exploration (IX) estimator, which yields clean bounds and avoids explicit uniform mixing [12, 17].

8.3 Regret guarantee

Theorem 3 (Structured regret)**.** *Against a non-anticipating adaptive adversary with losses in* $[0, 1]$ *satisfying the pooled-unseen symmetry, Algorithm 1 enjoys*

$$\mathbb{E}[\mathrm{Reg}_T] = O\left(\sqrt{T \log(T+1)}\right),$$

with the comparator being $\max_{b \in S_T \cup \{U\}} \sum_{t=1}^{T} \tilde{\ell}_t(b)$*, which equals* $\max_{j \in [K]} \sum_{t=1}^{T} \ell_t(j)$ *by Lemma 2.*

Proof sketch. By Lemma 2, the benchmark set has size $M \leq T+1$. The standard potential analysis of EXP3 with importance-weighted (or IX) estimates yields

$$\mathbb{E}[\mathrm{Reg}_T] \;\leq\; \frac{\log M}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}\left[\sum_{a \in \mathcal{A}_t} \frac{\mathrm{Var}(\widehat{\ell}_t(a) \mid H_{t-1})}{1}\right]^{1/2},$$

where the variance term is $O(1)$ per round for IX (or controlled via explicit mixing), and the prior term is $\log M \leq \log(T+1)$ due to average-weight initialization when new arms arrive. Optimizing $\eta \simeq \sqrt{\log M / T}$ gives the stated bound. See Auer et al. [2], Bubeck and Cesa-Bianchi [4] for EXP3 and Kocák et al. [12], Neu [17] for implicit exploration; adding experts over time with potential-preserving initialization is a standard device in specialist/growing-expert settings [16]. ☐

Algorithm 1 EXP3 (or EXP3-IX) with pooled-unseen reduction

Input: horizon T, learning rate $\eta > 0$, (optional) IX parameter $\gamma > 0$
Initialize $S_0 = \emptyset$, weights $w_1(a) = 1$ for $a \in \{U\}$; set $\mathcal{A}_1 = \{U\}$
for $t = 1$ to T **do**
  Form $\mathcal{A}_t = S_{t-1} \cup \{U\}$ and probabilities

$$p_t(a) = (1 - \mu_t) \frac{w_t(a)}{\sum_{b \in \mathcal{A}_t} w_t(b)} + \mu_t \cdot \frac{1}{|\mathcal{A}_t|}$$
(set $\mu_t = 0$ if using IX).

  Sample $A_t \sim p_t$, observe bandit loss $\ell_t(A_t)$.
  if $A_t = U$ **then** ▷ spawn a concrete arm $a^{\text{new}} \notin S_{t-1}$
    $S_t \leftarrow S_{t-1} \cup \{a^{\text{new}}\}$; define $\mathcal{A}_t \leftarrow S_t \cup \{U\}$
    Average-weight init: $w_t(a^{\text{new}}) \leftarrow \frac{1}{|\mathcal{A}_t|} \sum_{b \in \mathcal{A}_t} w_t(b)$
  else
    $S_t \leftarrow S_{t-1}$; $\mathcal{A}_{t+1} \leftarrow S_t \cup \{U\}$
  end if
  Form loss estimates for $a \in \mathcal{A}_t$:

$$\widehat{\ell}_t(a) = \begin{cases} \dfrac{\ell_t(A_t)}{p_t(A_t)} \mathbf{1}\{a = A_t\}, & \text{(standard EXP3)} \\ \dfrac{\ell_t(A_t)}{p_t(A_t) + \gamma} \mathbf{1}\{a = A_t\}, & \text{(EXP3-IX)} \end{cases}$$

  Update weights for $a \in \mathcal{A}_t$: $w_{t+1}(a) \leftarrow w_t(a) \exp(-\eta\, \widehat{\ell}_t(a))$.
end for

---

9 Conclusion

References

[1] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1638–1646, 2014. URL `https://proceedings.mlr.press/v32/agarwalb14.html`.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1): 48–77, 2002. doi: 10.1137/S0097539701398375. URL `https://epubs.siam.org/doi/10.1137/S0097539701398375`.

[3] J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrschein-*

lichkeitstheorie und Verwandte Gebiete, 47(2): 119–137, 1979. doi: 10.1007/BF00535278. URL https://link.springer.com/article/10.1007/BF00535278.

[4] S. Bubeck and N. 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.

[5] C. L. Canonne. A short note on an inequality between kl and tv. https://arxiv.org/abs/2202.07198, 2022.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921. URL https://cesa-bianchi.di.unimi.it/predbook/.

[7] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. URL https://www.schapire.net/papers/bandit-lin.pdf.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. URL https://dl.icdst.org/pdfs/files/aea72e61329cd4684709fa24f15ac098.pdf.

[9] I. Csiszár and J. Körner. Information theory: Coding theorems for discrete memoryless systems. *Cambridge University Press*, 2011. Contains Pinsker's inequality derivation.

[10] S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. arXiv preprint arXiv:1605.07416, 2016. URL https://arxiv.org/pdf/1605.07416.

[11] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2-3):245–272, 2010. doi: 10.1007/s10994-010-5178-7. URL https://link.springer.com/article/10.1007/s10994-010-5178-7.

[12] T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. URL https://papers.neurips.cc/paper_files/paper/2014/hash/8169cf3dc05090c7774c8dc38317c43d-Abstract.html.

[13] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2007. URL https://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandi

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

[15] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 2010. doi: 10.1145/1772690.1772758. URL https://arxiv.org/abs/1003.0146.

[16] J. Mourtada and O.-A. Maillard. Efficient tracking of a growing number of experts. In *Proceedings of the 30th Annual Conference on Learning Theory (COLT)*, pages 1325–1348, 2017. URL https://proceedings.mlr.press/v76/mourtada17a/mourtada17a.pdf.

[17] G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015. URL https://arxiv.org/abs/1506.03271.

[18] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. URL https://projecteuclid.org/journals/bulletin-of-the-american-mathematical-society/volume-58/issue-5/Some-aspects-of-the-sequential-design-of-experiments/bams/1183517370.full.

[19] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. doi: 10.1561/2200000018. URL https://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf.

[20] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/978-0-387-79052-7. URL https://link.springer.com/book/10.1007/b13794.