Reinforcement Learning

Lecture 8 : Bandit Identification

Rémy Degenne
(remy.degenne@inria.fr)

Centrale Lille, 2022/2023

Finding the best policy

Reinforcement Learning

- Interact with an unknown MDP
- Goal : Maximize the expected cumulative reward

Observations :

- There exists an optimal policy π^* independent of the starting state
- If an algorithm samples according to $\pi_t \approx \pi^*$, then it gets high expected cumulative reward

Results in reinforcement learning

For small MDPs with known dynamics :

Theorem

Value iteration converges in at most $\log\left(\frac{||T^\star(V_0)-V_0||_\infty}{\epsilon}\right)/\log(1/\gamma)$ iterations and outputs a policy π satisfying $||V^\pi - V^\star|| \leq \frac{\gamma\epsilon}{1-\gamma}$.

Theorem

Policy iteration terminates after a finite number of steps and outputs the optimal policy $\pi^\star$.

- No result on the actual sum of rewards obtained during learning.
- Only guaranty that we eventually approach π^*.
- Results only get worse for larger and unknown MDPs.

Regret minimization in bandits

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible
$\leftrightarrow$ minimizing the regret [Robbins, 1952]

$$\mathcal{R}_\nu(\mathcal{A}, T) := \underbrace{T\mu_\star}_{\substack{\text{sum of rewards of}\\ \text{an oracle strategy}\\ \text{always selecting } a_\star}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]}_{\substack{\text{sum of rewards of}\\ \text{the strategy} \mathcal{A}}}$$

Results :

- Lower bounds on the regret of consistent algorithms
- Algorithms with $O(\log T)$ regret upper bounds

Finding the best policy in bandits ?





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Maximize rewards $\leftrightarrow$ cure as many patients as possible

Alternative goal : identify as quickly as possible the best treatment (without trying to cure patients during the study)

Finding the best policy in bandits ?

$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t}$

Maximize rewards $\leftrightarrow$ cure as many patients as possible

Alternative goal : identify as quickly as possible the best treatment (without trying to cure patients during the study)

➜ Pure exploration, Best arm identification [Bubeck et al., 2011]

Best arm identification

Bandit interaction

At time t,

- choose an arm A_t
- observe a response $X_t \in \mathbb{R}$, sampled from distribution ν_{A_t}, with mean μ_{A_t}

Best arm identification goal : interact with the bandit for a while, then return the arm with highest mean.

Best arm identification

Bandit interaction

At time t,

- choose an arm A_t
- observe a response $X_t \in \mathbb{R}$, sampled from distribution ν_{A_t}, with mean μ_{A_t}

Best arm identification goal : interact with the bandit for a while, then return the arm with highest mean.

That is, find the best policy.

Goals : multiple objectives

Bandit interaction

At time t,

- choose an arm A_t
- observe a response $X_t \in \mathbb{R}$, sampled from distribution ν_{A_t}, with mean μ_{A_t}

Best arm identification goal : interact with the bandit for a while, then return the arm with highest mean.

Two goals

- Find the best arm with high probability
- Stop quickly

Let's formalize the problem

K arms with distributions $(\nu_1, \ldots, \nu_K)$, with means $(\mu_1, \ldots, \mu_K)$
At each time t, until the algorithm stops,

- choose an arm A_t
- observe a response $X_t \in \mathbb{R}$, sampled from distribution ν_{A_t}
- decide whether to stop or not

Let τ be the stopping time.
At τ, return $\hat{A}_\tau \in [K]$.

The algorithm makes a mistake if $\hat{A}_\tau \neq a^\star := \mathrm{argmax}_a \, \mu_a$.

Let's formalize the problem

K arms with distributions $(\nu_1, \ldots, \nu_K)$, with means $(\mu_1, \ldots, \mu_K)$
At each time t, until the algorithm stops,

- choose an arm A_t $\rightarrow$ sampling rule
- observe a response $X_t \in \mathbb{R}$, sampled from distribution ν_{A_t}
- decide whether to stop or not

Let τ be the stopping time. $\rightarrow$ stopping rule
At τ, return $\hat{A}_\tau \in [K]$. $\rightarrow$ recommendation rule

The algorithm makes a mistake if $\hat{A}_\tau \neq a^\star := \mathrm{argmax}_a \mu_a$.

Two problems

Two goals

- Find the best arm with high probability
- Stop quickly

Multiple objectives are hard to optimize simultaneously.

Solution : optimize one objective, under a constraint on the other.

- Fixed confidence identification :
  Optimize the stopping time of an algorithm, under a constraint on the probability of mistake
- Fixed budget identification :
  Optimize the probability of mistake after a given time

Outline

Fixed budget identification

Fixed budgete identification : minimize the probability of mistake after a given time.

Fixed Budget

Horizon T is known in advance, and the algorithm stops at $\tau = T$.

Goal : find an algorithm such that the probability of mistake $\mathbb{P}_\nu(\hat{A}_T \neq a^\star)$is as small as possible.

Simple algorithm : uniform sampling

Uniform sampling algorithm :

- sample all arms $\lfloor T/K \rfloor$ times $\rightarrow$ sampling rule
- return the best arm of the empirical mean vector $\hat{\mu}_T$ $\rightarrow$ recommendation rule

What is the probability of mistake ?

$$
\begin{aligned}
\mathbb{P}_\nu(\hat{A}_T \neq a^\star) &= \mathbb{P}_\nu(\operatorname*{argmax}_a \hat{\mu}_{T,a} \neq \operatorname*{argmax}_a \mu_a) \\
&= \mathbb{P}_\nu(\exists a \neq a^\star,\ \hat{\mu}_{T,a} > \hat{\mu}_{T,a^\star}) \\
&\leq \sum_{a \neq a^\star} \mathbb{P}_\nu(\hat{\mu}_{T,a} > \hat{\mu}_{T,a^\star}) \, .
\end{aligned}
$$

Concentration again

We need to bound $\mathbb{P}_\nu(\hat{\mu}_{T,a} > \hat{\mu}_{T,a^\star})$. Use a concentration inequality .

Hoeffding inequality

Z_i i.i.d. σ-sub-Gaussian random variables. For all $s \geq 1$

$$\mathbb{P}\left(\frac{Z_1 + \cdots + Z_s}{s} \geq \mu + x\right) \leq e^{-\frac{sx^2}{2\sigma^2}}$$

Both $\hat{\mu}_{T,a}$ and $\hat{\mu}_{T,a^\star}$ are averages of T/K i.i.d. random variables, with respective means μ_a and $\mu^\star$.

$$\begin{aligned}
&\mathbb{P}_\nu(\hat{\mu}_{T,a} > \hat{\mu}_{T,a^\star}) \\
&= \mathbb{P}_\nu(\hat{\mu}_{T,a} > \hat{\mu}_{T,a^\star}, \hat{\mu}_{T,a^\star} \leq \mu^\star - \frac{\Delta_a}{2}) + \mathbb{P}_\nu(\hat{\mu}_{T,a} > \hat{\mu}_{T,a^\star}, \hat{\mu}_{T,a^\star} > \mu^\star - \frac{\Delta_a}{2}) \\
&\leq \mathbb{P}_\nu(\hat{\mu}_{T,a^\star} \leq \mu^\star - \frac{\Delta_a}{2}) + \mathbb{P}_\nu(\hat{\mu}_{T,a} > \mu^\star - \frac{\Delta_a}{2}) \\
&= \mathbb{P}_\nu(\hat{\mu}_{T,a^\star} \leq \mu^\star - \frac{\Delta_a}{2}) + \mathbb{P}_\nu(\hat{\mu}_{T,a} > \mu_a + \frac{\Delta_a}{2}) \leq 2\exp\left(-\lfloor T/K \rfloor \frac{\Delta_a^2}{8\sigma^2}\right) .
\end{aligned}$$

Error probability of uniform sampling

Theorem

On the fixed budget best arm identification problem with budget T, uniform sampling has error probability

$$\mathbb{P}_\nu(\hat{A}_T \neq a^\star) \leq 2 \sum_{a \neq a^\star} \exp\left(-\lfloor T/K \rfloor \frac{\Delta_a^2}{8\sigma^2}\right)$$

This error probability is of order $\exp(-T\Delta_{\min}^2/(8K\sigma^2))$.

- Exponentially decreasing with T
- Rate of decrease of order $\Delta_{\min}^2/K$.

Oracle

Suppose we sample each arm n_a times, fixed in advance, not random, with $\sum_{a\in[K]} n_a = T$.
Return the best arm of the empirical mean vector $\hat{\mu}_T$.

Theorem

On the fixed budget best arm identification problem with budget T, that sampling scheme has error probability

$$\mathbb{P}_\nu(\hat{A}_T \neq a^\star) \leq K \sum_{a\neq a^\star} \exp\left(-n_{a^\star}\frac{\Delta_a^2}{8\sigma^2}\right) + \sum_{a\neq a^\star} \exp\left(-n_a\frac{\Delta_a^2}{8\sigma^2}\right)$$

We call static sampling oracle at μ the allocation $(n_a^\star)_{a\in[K]}$ which minimizes the probability of error.
It depends on μ (hence the name oracle) and verifies

$$\mathbb{P}_\nu(\hat{A}_T \neq a^\star) \leq K \exp\left(-\frac{T}{8\sigma^2\sum_a \frac{1}{\Delta_a^2}}\right) .$$

where $\Delta_{a^\star} = \min_{a\neq a^*} \Delta_a$.

Can we match the oracle ?

The static sampling oracle depends on the unknown μ, with $n^\star_a \approx \frac{1/\Delta_a^2}{\sum_b 1/\Delta_b^2}$.
Can we reach the same error probability without knowing μ ?

No, we can't [Carpentier and Locatelli, 2016]

Let $H(\mu) = \sum_a \frac{1}{\Delta_a^2}$. For any fixed budget identification algorithm, there exists a bandit problem with Gaussian arms with variance 1 such that

$$\mathbb{P}_\mu(\hat{A}_T \neq a^\star) \geq C_{K,T} \exp(-\frac{T}{H(\mu)\log K}) .$$

No algorithm can match the oracle rate of $\frac{T}{H(\mu)}$ everywhere.
But can we do almost as well ? Can we get $H(\mu) \log K$, since $H(\mu)$ is impossible ?

UCB-E

UCB for Exploration (UCB-E) [Audibert et al., 2010].

- Sample $A_t = \mathrm{argmax}_a \hat{\mu}_{t,a} + \sqrt{\frac{a}{N_{t,a}}}$.
- Recommend $\hat{A}_T = \mathrm{argmax}_a \hat{\mu}_{T,a}$.

Theorem

If UCB-E is run with parameter $0 < a \le \frac{25}{36}\frac{T-K}{H(\mu)}$, then it satisfies

$$\mathbb{P}_\mu(\hat{A}_T \neq a^\star) \le 2TK \exp(-\frac{2a}{25}) .$$

In particular for $a = \frac{25}{36}\frac{T-K}{H(\mu)}$, we have $\mathbb{P}_\mu(\hat{A}_T \neq a^\star) \le 2TK \exp(-\frac{T-K}{18H(\mu)})$.

Can match $T/H(\mu)$... if we know $H(\mu)$!

Successive Rejects

Idea : sample uniformly for a while, then reject the lowest arm. Sample the remaining arms uniformly, then reject the lowest, etc.

Successive Rejects [Audibert et al., 2010]

Let $\mathcal{A}_1 = [K]$, $\overline{\log}(K) = \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k}$, $n_0 = 0$ and for $k \in \{1, \ldots, K-1\}$,

$$n_k = \left\lceil \frac{T - K}{\overline{\log}(K)(K + 1 - k)} \right\rceil .$$

For each phase $k = 1, 2, \ldots, K_1$,

1. For each $a \in \mathcal{A}_k$, pull arm a for $n_k - n_{k-1}$ rounds.
2. Let $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{\mathrm{argmin}_{a \in \mathcal{A}_k} \hat{\mu}_{n_k, a}\}$.

Return the unique element of $\mathcal{A}_K$ as $\hat{A}_T$

Error probability of Successive Rejects

Theorem

The probability of error of successive rejects satisfies

$$\mathbb{P}_\mu(\hat{A}_T \neq a^\star) \leq K^2 \exp\left(-\frac{T-K}{\overline{\log}(K) H_2(\mu)}\right) ,$$

where $H_2(\mu) = \max_{k\in[K]} \frac{k}{\Delta_k^2}$.

$H_2(\mu) \leq H(\mu) \leq \log(K) H_2(\mu)$.

- Successive Rejects attains $T/(H(\mu)\log K)$ everywhere.

Open questions in fixed budget identification

- What is the complexity of parametric best arm identification ? (with Kullback-Leibler divergences and not gaps)
- What if the question is not to find the best arm, but something else about the distributions ? Lower bound, algorithms ?
- Can we have an algorithm that stops early if the problem is easy ?

Outline

Fixed confidence identification

Fixed confidence identification : Optimize the stopping time of an algorithm, under a constraint on the probability of mistake

δ-correct algorithm

An algorithm is said to be δ-correct on a set of bandit problems $\mathcal{D}$ if for all distribution tuples $\nu \in \mathcal{D}$,

$$\mathbb{P}_\nu(\hat{A}_\tau \neq a^\star) \leq \delta \, .$$

Goal : find a δ-correct algorithm such that the expected stopping time $\mathbb{E}_\nu[\tau]$ is as small as possible.

Variant : minimize $T_{\nu,\delta}$ such that with probability $1 - \delta$, the algorithm stops before $T_{\nu,\delta}$ and is correct.

Simple algorithm : uniform sampling

Idea : sample all arms in turn, until we can stop.

When is that ?

In addition to the sampling rule and the recommendation rule we need a stopping rule .

Stopping rule : confidence intervals

Concentration-based stopping rule :

- Maintain confidence intervals for the means of all arms
- Once the confidence interval of the best arm does not overlap with any other, stop

Recommendation rule : empirical best arm.

Suppose that with probability $1 - \delta$, the confidence intervals hold for all times.
Then with that probability : if the algorithm stops then the answer is correct.

- This is independent of the sampling rule !

Stopping rule : confidence intervals

Suppose that the arm distributions are σ^2-sub-Gaussian. Then

$$\mathbb{P}\left(\exists a, \exists t \in \mathbb{N},\ \hat{\mu}_{t,a} \notin \left[\mu_a - \sqrt{\frac{2\sigma^2 \log(\frac{2Kt^2}{\delta})}{N_{t,a}}}, \mu_a + \sqrt{\frac{2\sigma^2 \log(\frac{2Kt^2}{\delta})}{N_{t,a}}}\right]\right) \le \delta$$

Proof : Hoeffding's inequality, union bounds.

Uniform sampling

- Sample uniformly.
- Stop when the interval of the best arm does not overlap any other interval.
- Recommend that arm.

Theorem

With probability $1 - \delta$, that algorithm is correct and stops before

$$T_{\mu,\delta} := \inf \left\{ t \mid \sqrt{\frac{2\sigma^2 \log(Kt^2/\delta)}{t/K}} \leq \frac{\Delta_{\min}}{2} \right\} .$$

That is, $T_{\mu,\delta} \approx \frac{K}{\Delta_{\min}^2} 8\sigma^2 \log(K/\delta)$

Faster than uniform sampling ?

- Stop sampling arms that can be eliminated by another arm : Successive Elimination [Even-Dar et al., 2006]
  $T_{\mu,\delta} \approx (\sum_a \frac{1}{\Delta_a^2}) 8\sigma^2 \log(K/\delta)$

But what about the bad event of probability δ ?
If the best arm is eliminated, the algorithm might run for a very long time (see board).

- Sample the best arm and a well chosen challenger (LUCB [Kalyanakrishnan et al., 2012], Top Two algorithms [Russo, 2016, Jourdan et al., 2022])

We can get bounds on the expected stopping time $\mathbb{E}[\tau]$, also of order $(\sum_a \frac{1}{\Delta_a^2})\sigma^2 \log(1/\delta)$.

Towards optimality : lower bound

Our goal : get $\mathbb{E}[\tau]$ which is exactly as low as possible.

Lower bound

Any δ-correct algorithm on $\mathcal{D}$ verifies

$$\mathbb{E}_\nu[\tau] \max_{w \in \triangle_K} \inf_{\lambda \in \mathcal{D} : a^\star(\lambda) \neq a^\star(\nu)} \sum_a w_a \mathrm{KL}(\nu_a, \lambda_a) \geq \log \frac{1}{2.4\delta} .$$

Proof based on the chain rule and data processing inequality for the Kullback Leibler divergence.

GLRT stopping rule

$\mathcal{D}$ is a family of parametric distirbutions, parametrized by their means (technically we need a one-parameter exponential family).
$\mathrm{KL}(\mu_a, \lambda_a)$ for $\mu_a, \lambda_a \in \mathbb{R}$, denotes the KL between the corresponding distributions.
let $\hat{\mu}_t$ be the maximum likelihood estimator for the means at time t.

Lemma : LLR concentration

Under μ, with probability $1 - \delta$, for all $t \in \mathbb{N}$,

$$\sum_{s=1}^{t} \log \frac{d\mathbb{P}_{\hat{\mu}_{t,A_s}}}{d\mathbb{P}_{\mu_{A_s}}}(X_{s,A_s}) \le \log(\frac{t^2}{\delta}) \, .$$

Like we did with confidence intervals, we can get a stopping rule from this.

GLRT stopping rule

Lemma : LLR concentration

Under μ, with probability $1-\delta$, for all $t \in \mathbb{N}$,

$$\sum_{s=1}^{t} \log \frac{d\mathbb{P}_{\hat{\mu}_{t,A_s}}}{d\mathbb{P}_{\mu_{A_s}}}(X_{s,A_s}) \leq \log(\frac{t^2}{\delta}) \, .$$

Stop if

$$\inf_{\lambda \in \mathrm{alt}(\hat{\mu}_t)} \sum_{s=1}^{t} \log \frac{d\mathbb{P}_{\hat{\mu}_{t,A_s}}}{d\mathbb{P}_{\lambda_{A_s}}}(X_{s,A_s}) > \log(\frac{t^2}{\delta}) \, ,$$

where $\mathrm{alt}(\hat{\mu}_t) = \{\lambda \in \mathcal{D} \mid a^\star(\lambda) \neq a^\star(\hat{\mu}_t)\}$.
Return $\hat{A}_\tau = a^\star(\hat{\mu}_\tau)$.

$\rightarrow$ ensures δ-correct.

Why that likelihood ratio test ?

The expectation of a likelihood ratio is a KL :

$$\mathbb{E}_{X\sim\mu_a}[\log \frac{d\mathbb{P}_{\mu_a}}{d\mathbb{P}_{\lambda_a}}(X_a)] = \mathrm{KL}(\mu_a, \lambda_a)\,.$$

$$\mathbb{E}_\mu\left[\sum_{s=1}^t \log \frac{d\mathbb{P}_{\mu_{A_s}}}{d\mathbb{P}_{\lambda_{A_s}}}(X_{s,A_s})\right] = \mathbb{E}_\mu\left[\sum_{s=1}^t \mathrm{KL}(\mu_{A_s}, \lambda_{A_s})\right] = \sum_a \mathbb{E}[N_{t,a}]\mathrm{KL}(\mu_a, \lambda_a)$$

Suppose that we sampled each arm "tw_a times" and did not stop at t. Then

$$\begin{aligned}
\log(\frac{t^2}{\delta}) &\geq \inf_{\lambda\in \mathrm{alt}(\hat{\mu}_t)} \sum_{s=1}^t \log \frac{d\mathbb{P}_{\hat{\mu}_{t,A_s}}}{d\mathbb{P}_{\lambda_{A_s}}}(X_{s,A_s}) \\
&= t \inf_{\lambda\in \mathrm{alt}(\hat{\mu}_t)} \sum_a w_a \log \frac{d\mathbb{P}_{\hat{\mu}_{t,a}}}{d\mathbb{P}_{\lambda_a}}(\hat{\mu}_{t,a}) \\
&\approx t \inf_{\lambda\in \mathrm{alt}(\mu)} \sum_a w_a \mathrm{KL}(\mu_a, \lambda_a)\,.
\end{aligned}$$

Static proportions oracle

Suppose that we sampled each arm "tw_a times" (big enough for all a) and did not stop at t.

$$\log(\frac{t^2}{\delta}) \gtrsim t \inf_{\lambda \in \mathrm{alt}(\mu)} \sum_a w_a \mathrm{KL}(\mu_a, \lambda_a) \, .$$

Optimizing over w_a, we get something very close to the lower bound : for that optimal sampling (which depends on μ),

$$t \max_{w \in \triangle_K} \inf_{\lambda \in \mathrm{alt}(\mu)} \sum_a w_a \mathrm{KL}(\mu_a, \lambda_a) \lesssim \log(\frac{t^2}{\delta})$$

Track and Stop

Track and Stop

Sample every arm once, then at each time t until the algorithm stops,

1. Compute $\hat{w}_t^\star = \mathrm{argmax}_w \inf_{\lambda \in \mathrm{alt}(\hat{\mu}_t)} \sum_a w_a \log \frac{d\mathbb{P}_{\hat{\mu}_{t,a}}}{d\mathbb{P}_{\lambda_a}}(\hat{\mu}_{t,a})$
2. If there exists one arm with $N_{t,a} < \sqrt{t}$, pull it, (forced exploration) otherwise pull $A_t = \mathrm{argmin}_a N_{t,a} - t\hat{w}^\star_{t,a}$ (tracking)
3. Check the GLRT stopping rule

Recommend the empirical best arm

Theorem

Track-and-Stop is asymptotically optimal, that is

$$\lim_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau]}{\log(1/\delta)} \leq \frac{1}{\max_{w \in \triangle_K} \inf_{\lambda \in \mathrm{alt}(\mu)} \sum_a w_a \mathrm{KL}(\mu_a, \lambda_a)} .$$

Asymptotically optimal : upper bound identical to lower bound.

Limitations and improvements of TnS

Computing the argmax can be hard

- We can use an iterative method and do only one step at each time.

The forced exploration is harmful in practice

- We can introduce optimism to avoid it

Computing the argmin over the alternative could be hard in general identification problems.

Open problems in fixed confidence

- What is the complexity for δ not close to 0 ? Lower bounds and matching algorithms ?
- Can we have fixed confidence algorithms that we can choose to stop early, and still get error bounds ?

Reinforcement Learning

Ongoing research work :

- Can we get lower bounds on the time needed to find the best policy in RL ?
- Can we use the notion of alternative and apply methods like TnS ? (efficiently, preferably)

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010).
Best arm identification in multi-armed bandits.
Citeseer.

Bubeck, S., Munos, R., and Stoltz, G. (2011).
Pure exploration in finitely-armed and continuous-armed bandits.
Theoretical Computer Science, 412(19) :1832–1852.

Carpentier, A. and Locatelli, A. (2016).
Tight (lower) bounds for the fixed budget best arm identification bandit problem.
In *Conference on Learning Theory*, pages 590–604. PMLR.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006).
Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems.
Journal of machine learning research, 7(6).

Jourdan, M., Degenne, R., Baudry, D., de Heide, R., and Kaufmann, E. (2022).
Top two algorithms revisited.
arXiv preprint arXiv :2206.05979.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012).
Pac subset selection in stochastic multi-armed bandits.
In *ICML*, volume 12, pages 655–662.

Robbins, H. (1952).
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematical Society, 58(5) :527–535.

Russo, D. (2016).
Simple bayesian algorithms for best arm identification.
In *Conference on Learning Theory*, pages 1417–1418. PMLR.