HUMBOLDT-UNIVERSITÄT ZU BERLIN
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT
INSTITUT FÜR MATHEMATIK

Hierarchical optimistic optimization of $\mathcal{X}$-armed bandits

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

eingereicht von:	Philipp Trunschke
geboren am:	22.08.1992
geboren in:	Berlin
Gutachter/innen:	Prof. Dr. Markus Reiß Dr. Alexandra Carpentier
eingereicht am:	19. Oktober 2016

Contents

List of Symbols

$\{\Pi\}$	Characteristic function of Π. ($\{\Pi\}(x) = 1 \iff \Pi(x)$).
$X^{<\omega}$	Space of finite sequences over X.
ϵ	The empty sequence.
$q \succeq p$	p is a prefix of q. 17,
$\lvert p \rvert$	The length of the sequence p.
p^-	The sequence p with the last element removed.
$f(x) \in \mathcal{O}(g(x))$	f is bounded from above by g asymptotically.
$f(x) \in \tilde{\mathcal{O}}(g(x))$	$\exists k : f(x) \in \mathcal{O}\left(g(x) \ln^k g(x)\right)$.
$f(x) \in \Omega(g(x))$	f is bounded from below by g asymptotically.
$f(x) \in \Theta(g(x))$	$f(x) \in \mathcal{O}(g(x))$ and $f(x) \in \Omega(g(x))$.
$\mathcal{X}$	The search space. 2, 7, 17, 23,
M	The (noisy) function to optimize. 2, 7, 17, 23,
f	The mean payoff function $f(X) = \mathbb{E}[M_X]$. 2,
f^*, X^*	$f^* = \max f(\mathcal{X})$ and $X^* \in \operatorname{argmax} f(\mathcal{X})$. 3,
R_n, r_n	Cumulative and simple regret. 3,
$\mathcal{T}$	The set of indices. 7, 17,

$*, *[h]$	An optimal index. 7, 18,
$\mathscr{C}(X)$	The children of the nodes in X. 17,
$\mathscr{L}(X)$	The leafs of the subtree X of $\mathscr{T}$. 17,
P_t, $\mathscr{T}(n)$	See. 7, 18,
$\mathscr{X}_p, X_p$	Hierarchical partitioning of the set $\mathscr{X}$. 17,
Δ_p	See. 7, 18,
f_p^*, f_p°	See. 18,
Y_t	Reward of the t-th pull. 3,
$\hat{\mu}_p(t), T_p(t)$	See. 7, 23,
$U, B, \vec{B}$	See. 10, 18, 21, 23,
$\gamma_p(n,t), \gamma_p(n)$	See. 9,
δ_p, η_p	See. 19,
I, I_h	See. 19,
ℓ	A pre-metric. 31,
$\mathscr{B}_r(x)$	ℓ-Ball with center x and radius r. 33,
$\operatorname{diam}(X)$	ℓ-Diameter of X. 31,
ν, ρ	See. 31,
$\mathscr{X}_\epsilon$	$\mathscr{X}_\epsilon := \{X \in \mathscr{X} : f^* - f(X) \leq \epsilon\}$. 32,
q	Doubling constant of the pre-metric space $(\mathscr{X}, \ell)$. 34,
$\mathscr{N}, \mathscr{N}^{\text{pack}}, \mathscr{N}^{\text{cover}}$	Packing and covering numbers. 33,
d	The near-optimality dimension. 35,

To improve readability, we will interchangeably write g_x and $g(x)$ for any function g.

1. Introduction

1.1. Motivation

Maximizing a function M on a domain $\mathcal{X}$ is a fundamental problem in applied mathematics. In this thesis we will consider M to be stochastic. The function values may for example be generated by measurements or be solutions of stochastic algorithms that cannot be described deterministically. Such problems are ubiquitous to operations research and control. [Bubeck et al., 2011]

In adaptive clinical trials for example M may measure the effectiveness of a drug in relation to the severity of its side effects. The space of potential treatments $\mathcal{X}$ is possibly infinite and can be parametrized by the concentrations of the ingredients and the overall dose of the agent. The effectiveness $M(X)$ of a mixture X on a test subject is influenced by many unobservable factors that influence the outcome of the treatment (like lifestyle, genetic or environmental factors). In many relevant cases these effects can be modeled as a stochastic noise.

Another practical utilization is the optimization of marketing strategies (like the design of a website or the price of a product). For this, the testing period is divided into a fixed number of epochs in each of which the currently optimal strategy is implemented and the corresponding rewards are collected. The strategy to use in each epoch is chosen based on the results of the previous ones.[1]

Further potential applications include (without any claim of completeness)

- the stochastic optimization of networks (such as electrical, water, gas or traffic),
- minimizing the costs of transmitting information via a communication system in a noisy channel and
- calibrating the conditions of a chemical reaction [Bubeck et al., 2011].

These examples illustrate that an efficient algorithmic solution for such problems is desirable. In this thesis we will rederive several known algorithms which address this problem.

[1] Note, that the choice of the partition is crucial since a reward generated in over more than one epoch has to be disregarded.

1.2. Noisy functions and the bandit problem

Recall that a function $M : \mathcal{X} \to \mathbb{R}$ can be defined as an $\mathcal{X}$-indexed family $\{M_X\}_{X \in \mathcal{X}}$ of real numbers. In the same sense we can define a noisy function to be an $\mathcal{X}$-indexed family of probability distributions. An evaluation of M at X is a real-valued random variable $Y \sim M_X$, which is independent of all other evaluations. Consequently, a sequence of evaluations at a single point X gives a sequence of i.i.d. random variables $\{Y_i\}_{i \in \mathbb{N}}$.

Definition 1. *We call a pair $(\mathcal{X}, M)$ an **$\mathcal{X}$-armed bandit**[2], if $\mathcal{X}$ is measurable[3] and for each $X \in \mathcal{X}$*

- *M_X denotes the distribution of rewards associated with X*
- *M_X has support in $[0,1]$*
- *$f(X) = \mathbb{E}[M_X]$ attains a maximum on $\mathcal{X}$*

*If $|\mathcal{X}| = k$, then $(\mathcal{X}, M)$ is called a **k-armed bandit**.*

Remark 2. *Although we require the evaluations in each point to give a sequence of i.i.d. random variables, the following proofs will only require martingale difference sequences.*

Remark 3. *The nomenclature in this field is heavily influenced by the idea of a gambler playing on traditional slot machines. These machines were also called one-armed bandits because they were originally operated by a lever on the right side and were infamous for their ability to leave a gambler impoverished. [wik, b]*

*The machines are collected in the space $\mathcal{X}$ and their payoff distributions (which we assume not to change) are given by M. While playing, the gambler sequentially **pulls** the **arms** $X \in \mathcal{X}$ and collects **rewards** according to the arms payoff distribution M_X. Each pull advances her knowledge of the (empirical) distribution of that particular arm and she chooses the next arm to pull based on the sequence of pulls and rewards already obtained. [Bubeck et al., 2011, Auer et al., 2002, wik, a, Tekin, 2013]*

1.3. Policies and their assessment

The manner in which to choose the next arm to play, based on the previously selected arms and their rewards, is called an **allocation strategy** or **policy**. Formally, an allocation strategy is a sequence of measurable mappings

$$\phi_n : (\mathcal{X} \times [0,1])^{n-1} \to \mathcal{M}(\mathcal{X})$$

[2] This definition slightly varies from the definition given in [Bubeck et al., 2011].
[3] i.e. it is equipped with a σ-algebra

from the space of past observations $(\mathcal{X} \times [0,1])^{n-1}$ (equipped with the product σ-algebra) to the space $\mathcal{M}(\mathcal{X})$ of probability measures over $\mathcal{X}$.

Define Y_t to be the reward obtained in round t.

If $X_1, \ldots, X_{n-1}$ are the arms that have been pulled in the first $n-1$ rounds, then $\phi_n((X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1}))$ is a random variable that specifies which arm should be pulled in round n.

The measurability of ϕ_n ensures that this choice may be random, but can only depend on the information available in this round. If $\phi_n((X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1}))$ is a Dirac distribution for all n, then the strategy is called deterministic. Otherwise it is called randomized. [Bubeck et al., 2011]

Which policy to use depends on the objective of the gambler and a popular measure of its success in achieving this objective is the regret, that is the loss due to the fact that the globally optimal policy is not followed all the times. [Auer et al., 2002]

- If the objective is to find the machine $X \in \mathcal{X}$ which promises the highest expected reward, then the gambler can play n rounds until she is asked to make a recommendation $Z_n \in \mathcal{X}$. Her goal is to minimize the expectation of the **simple regret** [Bubeck et al., 2011]

$$r_n = f^* - f(Z_n)$$

  where $f^* = \max f(\mathcal{X})$.

- If, on the other hand, the objective is to maximize the expected value of the sum of rewards $\sum_{t=1}^{n} Y_t$, then her goal is to minimize the expectation of the **cumulative regret** [Bubeck et al., 2011]

$$\hat{R}_n = \sum_{t=1}^{n} (f^* - Y_t)$$

  By linearity and the tower property of the conditional expectation,
  this is equivalent to minimizing the expectation of the **cumulative pseudo-regret**[4] [Bubeck et al., 2011]

$$R_n = \sum_{t=1}^{n} (f^* - f(X_t))$$

Remark 4. *In both cases the policy needs to specify a sequence of pulls which balances between exploring the space to find new profitable arms and reexamining known arms.*

[4] $\mathbb{E}\left[\sum_{t=1}^{n}(f^* - \mathbb{E}[Y_t])\right] = \sum_{t=1}^{n}(f^* - \mathbb{E}[\mathbb{E}[Y_t|X_t]]) = \sum_{t=1}^{n}(f^* - \mathbb{E}[f(X_t)]) = \mathbb{E}\left[\sum_{t=1}^{n}(f^* - f(X_t))\right]$

In the second case, the empirically best arms even have to be played as often as possible. This dilemma is known as the exploration vs. exploitation trade-off in reinforcement learning. [Bubeck et al., 2011, Auer et al., 2002, wik, a, Tekin, 2013]

If the cumulative regret grows sublinear, then the gambler, in the long run, plays almost as well (up to a negligible term) as if she would follow the optimal strategy. In this case her policy is called Hannan consistent. [Bubeck et al., 2011]

Example 5. *Consider a two-armed bandit. A policy minimizing the simple regret may alternatingly pull each arm and return the arm with maximal empirical mean as the prediction. Whereas a policy minimizing the cumulative regret will rapidly focus on the empirically optimal arm.*

Example 6. *To illustrate the use of both regrets, consider a company developing a new drug which may cause substantial liver damage, if not measured out appropriately. The people in charge of this study decide to trial the drug in two stages. In phase one, it is applied to tissue samples and organoids and the company is interested in predicting the optimal concentrations and minimizing the regret of the individual predictions. When an adequate level of accuracy is achieved, the testing enters phase two in which the drug is administered to living test subjects. Now the main concern is to minimize the total number of test subjects which suffer liver damage.*

1.4. Setup

In the scope of this work we will focus on the minimization of the cumulative regret. The simple regret, however, can be bounded by the following lemma.

Lemma 7. *Let R_n be the cumulative pseudo-regret with respect to a given policy.*

- *Consider the recommendation $Z_n = X_{T_n}$, where X_t is the arm pulled in round t and T_n is drawn uniformly at random from $\{1, \dots, n\}$. Then $\mathbb{E}[r_n] \leq \frac{\mathbb{E}[R_n]}{n}$.*
- *Assume $\mathbb{E}[R_n] \in \Theta(g(n))$ for a non-negative, differentiable function g and consider the recommendation $Z_n = X_n$.*
  If g is concave, then $\mathbb{E}[r_n] \in \mathcal{O}(g'(n-1))$.
  If g is convex, then $\mathbb{E}[r_n] \in \mathcal{O}(g'(n))$.

Note that, since $\mathbb{E}[r_n]$ decreases, the first bound is useless if R_n grows superlinear.

Proof.

- $\mathbb{E}[r_n] = \mathbb{E}\left[f^* - f\left(X_{T_n}\right)\right] = f^* - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[f(X_n)] = \frac{\mathbb{E}[R_n]}{n}$[5]

[5] Adapted from [Bubeck et al., 2011]

- $r_n = R_n - R_{n-1} \in \Theta(g(n) - g(n-1))$
  Since g is differentiable, the mean value theorem ensures, that there exists an $x \in (n-1, n)$, such that $g(n) - g(n-1) = g'(x)$. Therefore $r_n \in \Theta(g'(x)) \subseteq \mathcal{O}(g'(x))$ and the assertion follows from concavity or convexity.

■

In the following section we will derive the HOO -algorithm. For this we will first consider two special cases.

In section 2.1 we assume that $\mathcal{X}$ is finite and that we can pull each arm at least once. Therefore we can find initial estimates for the expected rewards of each arm and use all subsequent pulls to reduce their error. The main result of this subsection will be the derivation of the algorithm UCB1 and its expected regret.

In section 2.2, $\mathcal{X}$ may be uncountable but we consider $M = f$ to be purely deterministic. Furthermore, we assume some knowledge about its continuity (Hoelder continuity for example). Then it suffices to pull a single arm to obtain knowledge about the maximal achievable function value in its neighborhood.
We introduce the policy DOO [6], which uses this information to determine an optimistic estimate for f and chooses the next arm based on this estimate.

In section 2.3 we will finally combine both results to obtain the HOO [7]-algorithm. For this we assume the mean f to satisfy the conditions for DOO and develop a modified version of this strategy. Since the evaluations are not deterministic, the necessary assumptions hold only in expectation and it is no longer guaranteed that the choices that were made by the policy are correct. We bound the probability of this happening with an argument similar to the one in section 2.1.

In the third section we will apply DOO and HOO to the special case that $\mathcal{X}$ is a pre-metric space.

[6] DOO is an acronym for deterministic optimistic optimization.
[7] HOO is an acronym for hierarchical optimistic optimization.

2. Derivation of the algorithm

2.1. UCB1

Let $((X_1,\dots,X_k),M)$ be a k-armed bandit and define the set of indices $\mathcal{T}=\{1,\dots,k\}$.
Let P_t be the index of the arm played in round t and define $\mathcal{T}(n)=(P_t)_{1\le t\le n}$.
Also define for any index p

$$T_p(n)=\sum_{t=1}^{n}\{P_t=p\} \qquad \text{and} \qquad \hat{\mu}_p(n)=\frac{1}{T_p(n)}\sum_{t=1}^{n}Y_t\{P_t=p\}$$

to be the number of times p has been played during the first n rounds and the empirical mean of the rewards obtained from those pulls.
We say that $p\in\mathcal{T}$ is **optimal** if $X_p=X^*$ and denote an optimal index by $*$. Define $\Delta_p:=f^*-f\left(X_p\right)$ and note that $\Delta_*=0$.

Note that instead of bounding $\mathbb{E}[R_n]$ directly, it is sufficient to bound $\mathbb{E}\left[T_p(n)\right]$ for suboptimal indices p, since

$$R_n=\sum_{t=1}^{n}\left(f^*-f\left(X_{P_t}\right)\right)=\sum_{t=1}^{n}\sum_{p=1}^{k}\Delta_p\{P_t=p\}=\sum_{\Delta_p\neq 0}\Delta_p T_p(n)$$

Probably the simplest policy is to fix an arm X_p which will be played in every round. This results in a linear growth of the cumulative regret with rate Δ_p, which is undesirable, if p is not optimal, since in this case, the regret is of the same order as the regret of the globally worst policy. A simple modification is the randomized policy ϵ-GREEDY .

Theorem 8 (Regret bound for ϵ-GREEDY)**.** *For the regret of* ϵ-GREEDY holds that

$$R_n\in\Theta(n)$$

Randomized policy: ϵ-GREEDY .
Parameters: $0 < \epsilon < 1$.
Loop: For each $n = 1, 2, \ldots$

- Let Q_n be any machine p that maximizes $\hat{\mu}_p(n)$.
- With probability $1-\epsilon$ play Q_n and with probability ϵ a random arm.

Proof. On the one hand

$$R_n = \sum_{t=1}^{n} \Delta_{P(t)} \leq n \max_{p \in \mathscr{T}} \Delta_p.$$

On the other hand, if m is the number of optimal arms, then

$$\begin{aligned}\mathbb{E}[R_n] &\geq \sum_{t=1}^{n} \left(\min_{\substack{p \in \mathscr{T} \\ \Delta_p \neq 0}} \Delta_p \right) \mathbb{P}\left[P_t \text{ is not optimal}^{1}\right] \\ &\geq n \left(\min_{\substack{p \in \mathscr{T} \\ \Delta_p \neq 0}} \Delta_p \right) \epsilon \frac{k-m}{k}.\end{aligned}$$

The last inequality holds, since by definition of ϵ-GREEDY ,

$$\begin{aligned}\mathbb{P}\left[P_t = p \mid \hat{\mu}_p(t) \text{ is maximal}^{2}\right] &= 1-\epsilon \\ \mathbb{P}\left[P_t = p \mid \hat{\mu}_q(t) \text{ is maximal}\right] &= \frac{\epsilon}{k}\end{aligned}$$

and consequently

$$\mathbb{P}\left[P_t = p,\ \hat{\mu}_p(t) \text{ is not max.}\right] = \sum_{q \neq p} \mathbb{P}\left[P_t = p \mid \hat{\mu}_q(t) \text{ is max.}\right] \mathbb{P}\left[\hat{\mu}_q(t) \text{ is max.}\right] \leq \frac{\epsilon}{k}$$

Therefore, the arm p will be played with probability

$$\begin{aligned}\mathbb{P}[P_t = p] &= \mathbb{P}\left[P_t = p \text{ and } \hat{\mu}_p(t) \text{ is not max.}\right] + \mathbb{P}\left[P_t = p \text{ and } \hat{\mu}_p(t) \text{ is max.}\right] \\ &\leq \frac{\epsilon}{k} + (1-\epsilon)\,\mathbb{P}\left[\hat{\mu}_p(t) \text{ is max.}\right]\end{aligned}$$

[1] id est $f\left(X_{P(t)}\right) \neq f^*$
[2] id est $\forall q : \hat{\mu}_q(t) \leq \hat{\mu}_p(t)$

and an optimal arm will be played with probability

$$\begin{aligned}\mathbb{P}[P_t \text{ is optimal}] &\leq \epsilon\frac{m}{k} + (1-\epsilon)\,\mathbb{P}\left[\hat{\mu}_*(t) \text{ is max.}^{3}\right] \\ &\leq \epsilon\frac{m}{k} + (1-\epsilon) \\ &= 1 - \epsilon\frac{k-m}{k}\end{aligned}$$

■

In the preceding proof we saw, that we can reduce the lower bound on R_n by reducing the exploration probability ϵ. This increases the sampling rate of the empirically optimal arm which in the long run will decrease the regret.

But although we can adjust it, ϵ-GREEDY still has a linear regret bound and it is natural to ask whether this is a consequence of poor algorithm design or a fundamental property of the problem itself.

Since the empirical mean of an arm becomes more accurate with every pull, we can dynamically modify its sampling rate to fit this accuracy.

An intuitive application of this concept is to modify ϵ-GREEDY by reducing ϵ with every step. This policy is called ϵ_n-GREEDY and [Auer et al., 2002] were able to prove, that its expected regret grows logarithmically.

An alternative approach is to assume

(Aγ) There exists a family $\left\{\gamma_p : \mathbb{N}^2 \to \mathbb{R}_+\right\}_{p\in\mathcal{T}}$ such that

- $\mathbb{P}\left[\hat{\mu}_p(n) \geq \mathbb{E}\left[\hat{\mu}_p(n) \mid \mathcal{T}(n)\right] + \gamma_p\left(n, T_p(n)\right)\right] \leq n^{-4}$ and $\mathbb{P}\left[\hat{\mu}_p(n) \leq \mathbb{E}\left[\hat{\mu}_p(n) \mid \mathcal{T}(n)\right] - \gamma_p\left(n, T_p(n)\right)\right] \leq n^{-4}$
- $\gamma_p(n,t)$ is increasing in n and decreasing in t
- $\lim_{t\to\infty}\gamma_p(n,t) = 0$.

The first condition of (Aγ) implies that $\mathbb{E}\left[\hat{\mu}_p(n) \mid \mathcal{T}(n)\right] \pm \gamma_p(n)$ give upper and lower confidence bounds on $\hat{\mu}_p(n)$ respectively. (For ease of notation we will write $\gamma_p(n) = \gamma_p\left(n, T_p(n)\right)$.) The second condition ensures, that the width of this bound decreases with every additional sample. And the third assumption guarantees, that indeed by collecting enough samples, we can make this bound arbitrarily tight.

Now we hope that we can sample the nodes often enough such that the confidence intervals $\left[\hat{\mu}_p(n) - \gamma_p(n), \hat{\mu}_p(n) + \gamma_p(n)\right]$ become pairwise disjoint.

[3] id est, one of the optimal arms has maximal empirical mean

Then with probability $1-n^{-4}$ the node with maximal $\hat{\mu}_p(n)$ will indeed be optimal.
Using this idea we can derive the deterministic policy UCB1 .

Theorem 9 (Regret bound for UCB1).

$$\mathbb{E}\left[T_p(n)\right] \leq \min\left\{t : 2\gamma_p(n,t) \leq \Delta_p\right\} + 4$$

Deterministic policy: UCB1 .
Initialization: Play each node once.
Loop: For each $n > k$ play the machine p that maximizes $\boldsymbol{B_p = \hat{\mu}_p(n) + \gamma_p(n)}$.

Proof. (adapted from [Auer et al., 2002])
Let $u \in \mathbb{R}$. Then after the k initialization rounds we have:

$$T_p(n) = 1 + \sum_{t=k+1}^{n} \{P_t = p\}$$

$$\{P_t = p\} = \left\{P_t = p, T_p(t-1) < u\right\} + \left\{P_t = p, T_p(t-1) \geq u\right\}$$

$$= 1 + \sum_{t=k+1}^{n} \left\{P_t = p,\ T_p(t-1) < u\right\} + \sum_{t=k+1}^{n} \left\{P_t = p,\ T_p(t-1) \geq u\right\}$$

$$\left\{P_t = p,\ T_p(t-1) < u\right\} = \sum_{s=1}^{u-1} \left\{P_t = p,\ T_p(t-1) = s\right\} \leq u-1$$

$$\leq 1 + (u-1) + \sum_{t=k}^{n-1} \left\{P_{t+1} = p,\ T_p(t) \geq u\right\}$$

$$P_{t+1} = p \text{ implies } B_*(t) \leq B_p(t)$$

$$\leq u + \sum_{t=1}^{\infty} \left\{B_*(t) \leq B_p(t),\ T_p(t) \geq u\right\}.$$

Now define the events

$$\Lambda \Leftrightarrow B_*(t) \leq B_p(t),\ T_p(t) \geq u$$

$$\Pi_1 \Leftrightarrow \hat{\mu}_p(t) + \gamma_p(t) \leq f(X_*)$$
$$\Pi_2 \Leftrightarrow \hat{\mu}_p(t) + \gamma_p(t) > f(X_*) \geq f\left(X_p\right) + 2\gamma_p(t)$$
$$\Pi_3 \Leftrightarrow f(X_*) < f\left(X_p\right) + 2\gamma_p(t)$$

By linearity and monotonicity of the expectation $\mathbb{E}\left[T_p(n)\right] \leq u + \sum_{t=1}^{\infty} \mathbb{P}[\Lambda]$. And since $\Pi_1 \vee \Pi_2 \vee \Pi_3 = 1$, we can split $\mathbb{P}[\Lambda] \leq \sum_{i=1}^{3} \mathbb{P}[\Lambda \wedge \Pi_i]$.

And by applying (Aγ) to each summand a straight forward computation shows

$$\begin{aligned}
\mathbb{P}[\Lambda \wedge \Pi_1] &\leq \mathbb{P}[\hat{\mu}_*(t) \leq f(X_*) - \gamma_*(t, T_*(t))] \leq t^{-4} \\
\mathbb{P}[\Lambda \wedge \Pi_2] &\leq \mathbb{P}\left[\hat{\mu}_p(t) \geq f\left(X_p\right) + \gamma_p\left(t, T_p(t)\right)\right] \leq t^{-4} \\
\mathbb{P}[\Lambda \wedge \Pi_3] &\leq \mathbb{P}\left[2\gamma_p\left(t, T_p(t)\right) > \Delta_p\right] \leq \mathbb{P}\left[2\gamma_p(n,u) > \Delta_p\right]
\end{aligned}$$

Finally note, that since $\lim_{u\to\infty} \gamma_p(n,u) = 0$, we can choose u such that $2\gamma_p(n,u) \leq \Delta_p$ and obtain the bound

$$\mathbb{E}\left[T_p(n)\right] \leq \min\left\{t : 2\gamma_p(n,t) \leq \Delta_p\right\} + \sum_{t=1}^{\infty} 2t^{-4} \leq \min\left\{t : 2\gamma_p(n,t) \leq \Delta_p\right\} + 4$$

■

Corollary 10. *Let $c_{t,s}$ be a sequence satisfying*

$$c_{t,s} \geq \sqrt{(8\ln t)/s} \tag{2.1}$$
$$r \leq s \Rightarrow c_{t,r} \geq c_{t,s} \tag{2.2}$$
$$\lim_{s\to\infty} c_{t,s} \to 0 \tag{2.3}$$

where we adapt the convention $c_{t,0} = +\infty$. Then $\gamma_p(t,s) = c_{t,s}$ satisfies (Aγ).

If in particular $c_{t,s} = \sqrt{(8\ln t)/s}$, then $\mathbb{E}[R_n] \leq \sum_{\Delta_p \neq 0} \frac{32(\ln n)}{\Delta_p} + 4\sum_{\Delta_p \neq 0} \Delta_p$.

Proof. Conditions two and three of (Aγ) follow directly from (2.2) and (2.3). So only the first condition remains to be shown. If $T_p(n) = 0$, then this condition holds trivially. So we can focus on the case $T_p(n) \geq 1$.

Define the stopping times $\widetilde{T}_j := \min\left\{t : T_p(t) = j\right\}$ and the random variables

$\widetilde{X}_j = X_{P(\widetilde{T}(j))}$ and $\widetilde{Y}_j = Y_{\widetilde{T}(j)}$. Then

$$\begin{aligned}
&\mathbb{P}\left[\hat{\mu}_p(n) \leq \mathbb{E}\left[\hat{\mu}_p(n) \mid \mathscr{T}(n)\right] - \gamma_p\left(n, T_p(n)\right) \text{ and } T_p(n) \geq 1\right] \\
&\mathbb{E}\left[\{P_t = p\} \mid \mathscr{T}(n)\right] = \{P_t = p\} \\
&\mathbb{E}\left[\{P_t \succeq p\} \mid \mathscr{T}(n)\right] = \{P_t \succeq p\} \\
&= \mathbb{P}\left[\sum_{j=1}^{T_p(n)} \widetilde{Y}_j \leq \sum_{j=1}^{T_p(n)} \mathbb{E}\left[\widetilde{Y}_j\right] - T_p(n)\,\gamma_p\left(n, T_p(n)\right) \text{ and } T_p(n) \geq 1\right] \\
&= \mathbb{P}\left[\sum_{j=1}^{T_p(n)} \left(\mathbb{E}\left[\widetilde{Y}_j\right] - \widetilde{Y}_j\right) \geq T_p(n)\,\gamma_p\left(n, T_p(n)\right) \text{ and } T_p(n) \geq 1\right] \\
&\leq \sum_{t=1}^{n} \mathbb{P}\left[\sum_{j=1}^{t} \left(\mathbb{E}\left[\widetilde{Y}_j\right] - \widetilde{Y}_j\right) \geq t\gamma_p(n,t)\right]
\end{aligned}$$

The last inequality follows, by union bound from $n \geq T_p(n) \geq 1$.

Direct computation shows, that

$$Z_t := \sum_{j=1}^{t} \left(\mathbb{E}\left[\widetilde{Y}_j\right] - \widetilde{Y}_j\right) = \sum_{j=1}^{t} \left(f\left(\widetilde{X}_j\right) - \widetilde{Y}_j\right)$$

is a martingale w.r.t. the filtration $\mathscr{F}_t = \sigma\left(\widetilde{X}_1, Z_1, \ldots, \widetilde{X}_t, Z_t, \widetilde{X}_{t+1}\right)$. And since by definition 1 the ranges of the martingale differences are bounded by 1, we can apply the Azuma-Hoeffding bound:

$$\begin{aligned}
\mathbb{P}\left[\sum_{j=1}^{t} \left(\mathbb{E}\left[\widetilde{Y}_j\right] - \widetilde{Y}_j\right) \geq t\gamma_p(n,t)\right] &\leq \exp\left(-\frac{\left(t\gamma_p(n,t)\right)^2}{2t}\right) \\
&\leq \exp\left(-\frac{t}{2}\left(\sqrt{(8\ln n)/t}\right)^2\right) \\
&= n^{-4}
\end{aligned}$$

Now let $c_{t,s} = \sqrt{\frac{8\ln t}{s}}$, then

$$2c_{n,t} \leq \Delta_p \iff 32(\ln n)/t \leq \Delta_p^2 \iff t \geq \frac{32(\ln n)}{\Delta_p^2}$$

Therefore $\mathbb{E}\left[T_p(n)\right] \leq \frac{32(\ln n)}{\Delta_p^2} + 4$ and

$$\mathbb{E}[R_n] = \sum_{\Delta_p \neq 0} \Delta_p \mathbb{E}\left[T_p(n)\right] \leq \sum_{\Delta_p \neq 0} \frac{32(\ln n)}{\Delta_p} + 4 \sum_{\Delta_p \neq 0} \Delta_p$$

■

Example 11. *Define $\widetilde{Y}_j$ as in the preceding proof and $\widetilde{\mu}_p(t) = \frac{1}{t}\sum_{j=1}^{t} \widetilde{Y}_j$.* [4]

- *The only place we needed M to have compact support was the application of the Azuma-Hoeffding-bound in the proof of the preceding corollary. Let us now consider $M_{X_p} = \mathcal{N}\left(f\left(X_p\right), \sigma_p^2\right)$.* [5]
  Then $\widetilde{\mu}_p(t) - f\left(X_p\right) \sim \mathcal{N}\left(0, \frac{\sigma_p^2}{t}\right)$ for $t > 0$ and we know that

$$\mathbb{P}\left[\widetilde{\mu}_p(t) \leq f\left(X_p\right) - \gamma_p(n,t)\right] = F_{0,\sigma_p^2/t}\left(-\gamma_p(n,t)\right) \overset{!}{\leq} n^{-4}$$

  Therefore we can choose

$$\gamma_p(n;t) \geq -F^{-1}_{0,\sigma_p^2/t}\left(n^{-4}\right) = -\frac{\sigma_p}{\sqrt{t}}\operatorname{probit}\left(n^{-4}\right) \approx [6] \frac{\sigma_p}{\sqrt{t}}\sqrt{\frac{\pi}{8}}\ln\left(t^4 - 1\right)$$

  It is easy to see that this choice satisfies $(A\gamma)$. Finally, note that

$$\begin{aligned} 2\gamma_p(n,t) \leq \Delta_p &\Leftarrow F^{-1}\left(n^{-4}\right) \geq -\frac{\Delta_p}{2} \\ &\Leftrightarrow n^{-4} \geq \frac{1}{2}\left(1 + \operatorname{erf}\left(-c\sqrt{t}\right)\right) \\ &\Leftarrow t \geq c^{-2}\left(\operatorname{erf}^{-1}\left(2n^{-4} - 1\right)\right)^2 \end{aligned}$$

  where $c = \frac{\Delta_p}{\sqrt{8}\sigma_p}$. And therefore $R_n \in \mathcal{O}\left(c^{-2}\left(\operatorname{erf}^{-1}\left(2n^{-4} - 1\right)\right)^2\right)$.

[4] Recall that $\widetilde{Y}_j$ depends implicitly on p.

[5] In this example $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the normal distribution with mean μ and variance σ^2. Its CDF is given by

$$F_{\mu,\sigma^2}(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right)$$

and its quantile function by

$$F^{-1}_{\mu,\sigma^2}(p) = \mu + \sigma \operatorname{probit}(p)$$

where $\operatorname{probit}(p) = \sqrt{2}\operatorname{erf}^{-1}(2p-1)$.

[6] We use the Logit function as a simple approximation for the probit function.

The growth of R_n in comparison to $(\ln n)$ is illustrated in figure 2.1. There we can see, that the additional generality is paid for by a worse bound on the regret.

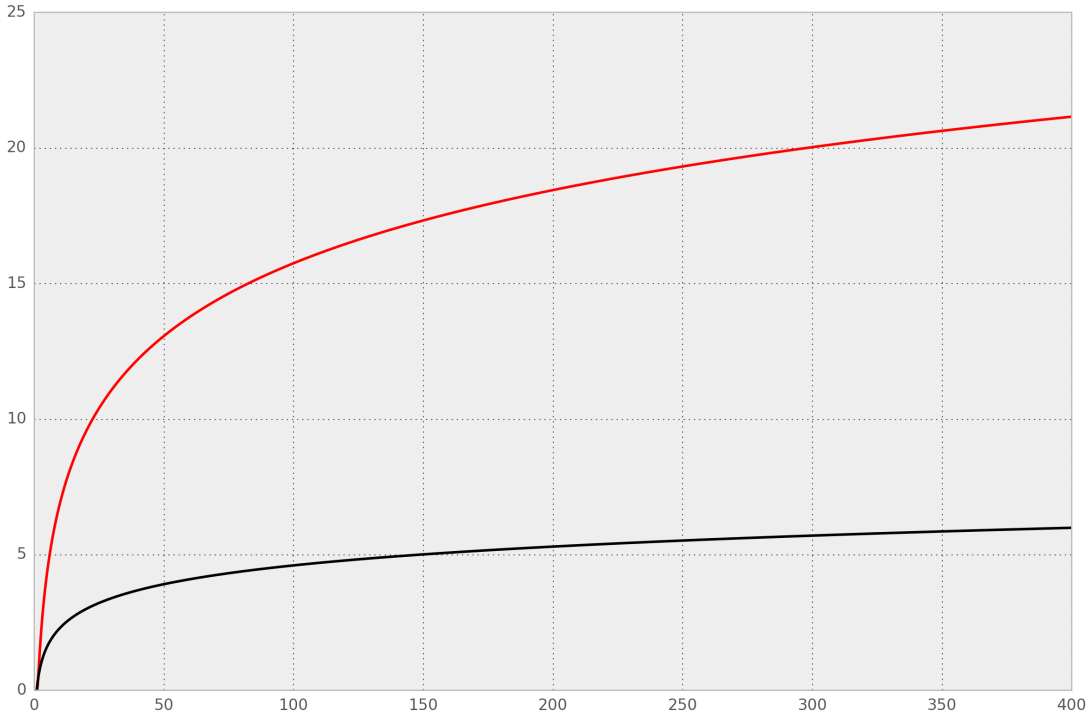


Figure 2.1.: Plot of $\left(\operatorname{erf}^{-1}\left(2n^{-4}-1\right)\right)^2$ (red) and $(\ln n)$ for $n \in [1, 400]$.

- *The more we know about the distributions M, the better we can choose γ_p and achieve a potentially smaller bound for $T_p(n)$.*

  Take for example $M\left(X_p\right) = \mathcal{U}\left(f\left(X_p\right) - \sigma_p, f\left(X_p\right) + \sigma_p\right)$.[7]
  Then $\widetilde{\mu}_p(t) - f\left(X_p\right) \sim \text{Bates}\left(-\sigma_p, \sigma_p;\ t\right)$ for $t > 0$ and we could again consider the choice $\gamma_p(n,t) = -F^{-1}\left(n^{-4}\right)$.[8] *Again*

$$2\gamma_p(n,t) \leq \Delta_p \Leftarrow n^{-4} \geq F\left(-\frac{\Delta_p}{2}\right)$$

  But now, if $\frac{\Delta_p}{2} \geq \sigma_p$, then $F\left(-\frac{\Delta_p}{2}\right) = 0$ for all t and consequently $R_n \in \mathcal{O}(1)$.

[7] In this example $\mathcal{U}(a,b)$ denotes the uniform distribution on $[a,b]$.
[8] Now F denotes the CDF of the Bates distribution.

Remark 12. *If γ and $\overline{\gamma}$ are two families satisfying $(A\gamma)$ and if γ is such that $\underline{\gamma} \leq \gamma \leq \overline{\gamma}$, then γ satisfies $(\overline{A}\gamma)$ as well and*

$$\min\left\{t : 2\underline{\gamma}_p(n;t) \leq \Delta_p\right\} \leq \min\left\{t : 2\gamma_p(n;t) \leq \Delta_p\right\} \leq \min\left\{t : 2\overline{\gamma}_p(n;t) \leq \Delta_p\right\}$$

Remark 13. *Originally [Auer et al., 2002] proved theorem 9 by using the Chernoff-Hoeffding-bound for empirical means instead of the more general Azuma-Hoeffding-bound. This allowed them to choose $\gamma_p(t,s) = \sqrt{\frac{2\ln t}{s}}$ and obtain a slightly tighter bound on $\mathbb{E}\left[T_p(n)\right]$.*

2.2. DOO

In this chapter we consider a bandit $(\mathscr{X}, M)$ with infinite $\mathscr{X}$ and deterministic M_X.

For $k \in \mathbb{N}_{\geq 1}$ define $\mathscr{T} := \{1, \dots, k\}^{<\omega}$ and endow it with the semi-order $\preceq$ where

$$p \preceq q \Leftrightarrow p \text{ is a prefix of } q$$

With this definition $\mathscr{T}$ becomes a non-degenerate infinite k-ary tree. In this context we call $p \in \mathscr{T}$ a **path** and say, that it lies at depth $|p|$ in $\mathscr{T}$. If $p \succeq q$ we will also say, that q lies on the path to p.

Define for any set of nodes $X \subseteq \mathscr{T}$ the set of their children as

$$\mathscr{C}(X) := \left\{p \in \mathscr{T} : p^- \in X\right\}$$

and for any subtree $X \subseteq \mathscr{T}$ define the set of its leafs as

$$\mathscr{L}(X) := \{x \in X : \mathscr{C}(\{x\}) \cap X = \emptyset\}$$

Now we assume the existence of a hierarchical partitioning of $\mathscr{X}$:

(A$\mathscr{X}$) There exists a family of subsets $\mathscr{X}_p \subseteq \mathscr{X}$, indexed by $p \in \mathscr{T}$, such that

- $\mathscr{X}_\epsilon = \mathscr{X}$ (where ϵ denotes the empty string)
- $\mathscr{X}_p \subseteq \bigcup \left\{\mathscr{X}_q : q \in \mathscr{C}(\{p\})\right\}$

and a family $X_p \in \mathscr{X}_p$.

Example 14. *Let $\mathscr{T}$ be an infinite complete binary tree ($k = 2$ and $\mathscr{T} = \{1,2\}^{<\omega}$) and define*

$$\mathscr{X}_\epsilon = [-1, 1]$$

$$\mathscr{X}_{s0} = [\min \mathscr{X}_s, X_s]$$
$$\mathscr{X}_{s1} = [X_s, \max \mathscr{X}_s]$$

where $X_s = \frac{1}{2}(\min \mathscr{X}_s + \max \mathscr{X}_s)$.

Remark 15. *Instead of requiring $\mathscr{X}_p \subseteq \bigcup \left\{\mathscr{X}_q : q \in \mathscr{C}(\{p\})\right\}$ in (A$\mathscr{X}$) it would be sufficient to ensure $X^* \in \mathscr{X}_s \Rightarrow X^* \in \bigcup_i \mathscr{X}_{si}$.*

2. Derivation of the algorithm

We now define

$$f_p^* := \sup f\left(\mathcal{X}_p\right), \qquad f_p^\circ := \inf f\left(\mathcal{X}_p\right) \quad \text{and} \quad \Delta_p := f^* - f_p^\circ.$$

Lemma 16. $p \succeq q \Rightarrow \Delta_p \leq \Delta_q$

Proof. $p \succeq q \Rightarrow \mathcal{X}_p \subseteq \mathcal{X}_q \Rightarrow \Delta_p \leq \Delta_q$ ■

We call a node $p \in \mathcal{T}$ optimal, if $X^* \in \mathcal{X}_p$ and as in [Valko et al., 2013], we define a sequence of optimal nodes $*[h]$ such that $|*[h]| = h$ and $*[h+1]^- = *[h]$.

As in the preceding section let P_t denote the path to the arm played in round t and define $\mathcal{T}(n) := (P_t)_{1 \leq t \leq n}$. We will interpret $\mathcal{T}(n)$ as a subtree $\mathcal{T}(n) \subseteq \mathcal{T}$ of $\mathcal{T}$.

Lemma 17. *If there exists a tree* $\left\{\delta_p\right\}_{p \in \mathcal{T}}$ *such that* $\delta_{*[h]} \geq \Delta_{*[h]}$ *and* $\lim_{|p| \to \infty} \delta_p = 0$ *then, if the following policy is applied*

$$B_{Q(n)} \geq f^* \qquad \textit{and} \qquad \lim_{n \to \infty} f\left(X_{P(n)}\right) = f^*$$

Deterministic policy: DOO .
Parameters: f and the trees X and δ.
Initialization: Define $\boldsymbol{P_0 = \epsilon}$.
Loop: For $n \in k\mathbb{N}$[9] sample $\boldsymbol{\{P_{n+1}, \ldots, P_{n+k}\} = \mathscr{C}\left(\{Q_n\}\right)}$ where
$\boldsymbol{Q_n \in \operatorname*{argmax}_{p \in \mathcal{L}\mathcal{T}(n)} B_p}$ and $\boldsymbol{B_p := f\left(X_p\right) + \delta_p}$.
Recommendation: $Z_n = \operatorname*{argmax}_{p \in \mathcal{T}(n)} f\left(X_p\right)$.

Remark 18. *To understand, what* DOO actually does in step n lets take a look on the sequence P_n. For $n = 0$ DOO defines $P_0 = \epsilon$ in the initialization.

For $n > 0$, DOO proceeds in two stages.
In the first stage it selects a leaf $Q_{n-1} \in \mathcal{T}(n-1)$ that has maximal B-value. Then, for the next k rounds, DOO samples each child of Q_{n-1} before going back to stage one.

Proof.

$$\lim_{n \to \infty} B_{Q(n)} = \lim_{n \to \infty} f\left(X_{Q(n)}\right) + \lim_{n \to \infty} \delta_{Q(n)} = \lim_{n \to \infty} f\left(X_{Q(n)}\right)$$

[9] $k\mathbb{N} := \{kn : n \in \mathbb{N}\}$

For every n there has to exist an optimal node $*[h] \in \mathscr{LT}(n)$.
By assumption $B_{*[h]} \geq f^*$ and by definition of DOO $B_{Q(n)} \geq B_{*[h]} \geq f^*$.
This implies $\lim_{n\to\infty} B_{Q(n)} = \lim_{n\to\infty} f\left(X_{Q(n)}\right) \geq f^*$
and $f(X) \leq f^*$ implies $\lim_{n\to\infty} f\left(X_{Q(n)}\right) \leq f^*$.
Finally lemma 16 implies $0 \leq \lim_{n\to\infty} \Delta_{P(n)} \leq \lim_{n\to\infty} \Delta_{Q(n)} = 0$. ■

Remark 19.

- *In lemma 17 it is sufficient to demand $f\left(X_p\right) + \delta_p \geq f^*$ instead of $\delta_{*[h]} \geq \Delta_{*[h]}$.*
- *If, however, δ is chosen as in lemma 17, then the same argument can be applied to $B_p = f_p + \delta_p$ for any $f_p \in \left[f_p^\circ, f_p^*\right]$.*

If δ is a tree satisfying the assumptions of lemma 17, then

$$\delta'_p := \max_{\substack{q\in\mathscr{T}\\|q|=|p|}} \sup_{r\succeq q} \delta_r$$

satisfies them as well. Therefore, we will assume in the following that

(Aδ) $\qquad \exists \delta_h \overset{h\to\infty}{\longrightarrow} 0 \text{ decreasing} : \delta_h \geq \Delta_{*[h]}$[11]

Whenever the children of a node $Q(n)$ are being sampled we say, that $Q(n)$ is being expanded. In lemma 17 we seen that $B_{Q(n)} \geq f^*$. That means, that a node p such that $B_p < f^*$ will never be expanded and that DOO only expands nodes of the set

$$I = \left\{p : f_p^* + \delta_p \geq f^*\right\}$$

To bound the number of nodes that may be expanded by DOO we define $I_h = \{p \in I : |p| = h\}$ and

(Dη) $\qquad \eta \in \mathbb{R}^{\mathbb{N}} \text{ such that } \eta_h \geq |I_h|$

Theorem 20 (Regret bound for DOO). *Let $h(n) = \inf\left\{h : \sum_{l=0}^{h} \eta_l \geq n\right\}$, then*

$$r_n \leq \delta_{h(n)} \tag{2.4}$$

$$R_n \leq k \sum_{l=0}^{h(n)} \eta_l \delta_l \tag{2.5}$$

[11] By an abuse of notation we will interchangeably write $\delta_p = \delta_{|p|}$.

Proof. (adapted from [Munos, 2011])
(r_n) Let $p_{\max}$ be the deepest node that has been expanded up to round n. Then by definition of Z_n and since by lemma 17 $B_{p_{\max}} \geq f^*$

$$f(Z_n) \geq f\left(X_{p_{\max}}\right) \geq f^* - \delta_{p_{\max}}$$

By (Aδ) we can bound $f(Z_n) \geq f^* - \delta_p$ for any path p with $|p| \leq |p_{\max}|$. Since every node is sampled just once, $\mathscr{T}(n)$ has to exceed a minimal height. This height is attained, when as much of the lower levels as possible are fully expanded.
The number of nodes which are being sampled at height h is bounded by $k\eta_{h-1}$. Therefore the height of $\mathscr{T}(n)$ has to be at least $h(n) := \inf\left\{h : \sum_{l=0}^{h} k\eta_l \geq n\right\}$. Consequently

$$f(Z_n) \geq f^* - \delta_{h(n)}$$

(R_n)

$$\begin{aligned} R_n &= \sum_{t=1}^{n}\left(f^* - f\left(X_{P(n)}\right)\right) \leq \sum_{t=1}^{n} \Delta_{P(n)} \\ &\leq \sum_{t=1}^{n} \Delta_{Q(n)} && \text{lemma 16} \\ &\leq \sum_{t=1}^{n} \delta_{Q(n)} && Q_n \in I \end{aligned}$$

Since $\delta_p = \delta_{|p|}$ is decreasing with $|p|$ we can bound this sum by assuming that we are expanding all nodes in I_h before expanding any node in I_{h+1}. The number of levels we can expand in this way is again bounded by $h(n)$. And since for each node we expand we have to sample its k children we obtain the bound

$$R_n \leq \sum_{l=0}^{h(n)} k\eta_l \delta_l.$$

■

Remark 21. *It is advantageous to find δ as small as possible, since $\underline{\delta} \leq \delta$ implies*

$$\left\{p : f\left(X_p\right) + \underline{\delta} \geq f^*\right\} \subseteq \left\{p : f\left(X_p\right) + \delta \geq f^*\right\}$$

Therefore $\underline{\eta} \leq \eta$ and $\underline{R}_n \leq R_n$. (where $\underline{\eta}$ and $\underline{R_n}$ are defined as expected)

Example 22. *If $\mathscr{X}$ is defined as in example 14 and $f(X) = 1 - |X|^\alpha$ (for $\alpha \geq 0$), then*

$$f^* - \min f\left(\mathscr{X}_{*[h]}\right) = \max\left\{|X|^\alpha : X \in \mathscr{X}_{*[h]}\right\} \leq 2^{-\alpha h}.$$

If we choose $\delta = 2^{-\alpha h}$*, then*

$$\begin{aligned} |I_h| &= \left|\left\{p : |p| = h,\ f\left(X_p\right) \geq 1 - 2^{-\alpha h}\right\}\right| \\ &= \left|\left\{p : |p| = h,\ \left|X_p\right|^{\alpha} \leq 2^{-\alpha h}\right\}\right| \\ &\leq 2 \end{aligned}$$

Therefore $h(n) = \inf\{h : h2 \geq n\} = \frac{n}{2}$ *and*

$$\begin{aligned} r_n &\leq 2^{-\alpha n/2} \\ R_n &\leq 2\sum_{l=0}^{n/2} 2^{-\alpha l} = 2\frac{1 - 2^{-\alpha(n/2+1)}}{1 - 2^{-\alpha}} \end{aligned}$$

Remark 23. *In the policy* DOO , only the values associated to the nodes in $\mathscr{LT}(n)$ have to be memorized.

Remark 24. *We can implement an alternative version of* DOO by

Deterministic policy: DOO2 .
Parameters: f and the trees X and δ.
Initialization: Define $B_p = +\infty$ for all $p \in \mathscr{T}$.
Loop: For each $n = 1, 2, \ldots$

- Expand $P_n \in \operatorname{argmax}\left\{\vec{B}_p : p \in \mathscr{C}(\mathscr{T}(n))\right\}$
  where $\vec{B}_p = \left(B_{p_0}, B_{p_0 p_1}, \ldots, B_p\right)$ are ordered lexicographically.[12]
- Compute $U_{P(n)} = f\left(X_{P(n)}\right) + \delta_{P(n)}$.
- Update $B_p = \min\left\{U_p, \max_{\mathscr{C}(p)}\left\{B_q\right\}\right\}$.

Recommendation: $Z_n = \operatorname*{argmax}_{p \in \mathscr{T}(n)} f\left(X_p\right)$.

To clearify the definition of the sequence P_t lets first take a look on the definition of the B-values for the nodes $p \in \mathscr{T}(n)$.

We can argue by induction, that B_p always is an optimistic estimate for the maximal value of f on $\mathscr{X}_p$.

If not all children of p have been sampled yet, then $B_p = U_p$, which indeed is such an optimistic estimate. Note that $B_p = U_p$ also implies that DOO2 , like DOO , first samples all children of a node before expanding the next.

If all children $q \in \mathscr{C}(\{p\})$ have been sampled then by induction hypothesis each B_q is an optimistic bound for f on $\mathscr{X}_q$. Therefore both U_p and $\max_{q \in \mathscr{C}(\{p\})} B_q$ are bounds for f on $\mathscr{X}_p$ and since B_p is their minimum, it must be too.

Now lets return to the sequence P_t. To find the node to sample next, DOO2 builds the path P_t step by step. Starting at the root $p = \epsilon$ in each step it compares the B-values of the children of p and redefines p to be the path to that child with the highest B-value until an unexplored node $p \notin \mathscr{T}(n)$ is reached.

Note that the nessecary conditions for the proof of the regret bound are satisfied.

Since both algorithms have the same regret bound it might seem far-fetched to compare all B-values on the paths to the leaves instead of just the values at the leaves themselves. Especially since this has the disadvantage, that we have to keep the whole tree $\mathscr{T}(n)$ in memory instead of just the leafs $\mathscr{L}\mathscr{T}(n)$. However, if the tree is not unary, then $|\mathscr{T}(n)|$ and $|\mathscr{L}\mathscr{T}(n)|$ differ on average only by a factor. In a balanced k-ary tree, for example, the number of leaves $|\mathscr{L}\mathscr{T}(n)|$ is of order $k^{(\log_k n)-1} = n/k$. Also, in HOO the U-values will change every round and we will need to reconsider each choice.

2.3. HOO

In this section we consider $(\mathcal{X}, M)$ to be an arbitrary $\mathcal{X}$-armed bandit. That is, $\mathcal{X}$ may be infinite and M may not be deterministic.
We define $\mathcal{T}, P_t$ and $\mathcal{T}(n)$ as in the preceding section and assume that $(A\mathcal{X})$, $(A\delta)$ and $(D\eta)$ hold for f. Analogous to UCB1 2.1 we define

$$T_p(n) = \sum_{t=1}^{n} \{P_t \succeq p\} \qquad \text{and} \qquad \hat{\mu}_p(n) = \frac{1}{T_p(n)} \sum_{t=1}^{n} Y_t \{P_t \succeq p\}$$

Our ansatz is to apply DOO to the function f. But since we cannot sample f directly, we have to replace $B_p = f\left(X_p\right) + \delta_p$ by a suitable alternative.

Therefore we assume $(A\gamma)$ and substitute $f\left(X_p\right)$ by $\hat{\mu}_p(n) + \gamma_p(n)$. $(A\gamma)$ allows us to estimate with probability $1 - n^{-4}$, that $B_{*[h]} \geq f^*$. But since $\left(\hat{\mu}_p(n) + \gamma_p(n)\right)$ changes every round, we need to use the implementation DOO2 .

Theorem 25 (Regret bound for HOO). *Let δ, η be chosen such that $(A\delta)$ and $(D\eta)$ are satisfied for f and choose $\gamma_p(n,t) \geq \sqrt{\frac{8 \ln n}{t}}$ such that $(A\gamma)$ is satisfied.*
Then for every $H \in \mathbb{R}_+$ the cumulative regret is bounded by

$$\mathbb{E}[R_n] \leq n\delta_H + \sum_{h=0}^{H-1} \eta_h \delta_h + k \sum_{h=1}^{H} \eta_{h-1} \delta_{h-1} \tau_h(n)$$

where $\tau_h(n) = \max_{|p|=h} \tau_p(n)$ and $\tau_p(n) \leq \frac{32 \ln n}{\left(\Delta_p - \delta_p\right)^2} + 4$ is an upper bound for $\mathbb{E}\left[T_p(n)\right]$ for $p \notin I_h$.

Deterministic policy: HOO .
Parameters: M and the trees X, δ and γ.
Initialization: Define $B_p = +\infty$ for all $p \in \mathcal{T}$.
Loop: For each $n = 1, 2, \ldots$

- Expand $P_n \in \operatorname{argmax}\left\{\vec{B}_p : p \in \mathcal{C}(\mathcal{T}(n))\right\}$
  where $\vec{B}_p = \left(B_{p_0}, B_{p_0 p_1}, \ldots, B_p\right)$ are ordered lexicographically.
- Update $U_p = \hat{\mu}_p(n) + \gamma_p(n) + \delta_p$ and $B_p = \min\left\{U_p, \max_{\mathcal{C}(p)}\{B_q\}\right\}$
  for all $p \in \mathcal{T}(n)$.

Proof. (Adapted from [Bubeck et al., 2011])

Inspired by the proof of theorem 20, we want to split the tree $\mathcal{T}$ into the tree I and its compliment. Since it is no longer guarantueed that $P_t \in I$, we have to bound $\mathbb{E}\left[T_q(n)\right]$ for $q \in I^c$. For this, each level h of the tree can be considered as a k^h-armed bandit and we can draw on what we learned in the proof of theorem 9 to find the bound $\mathbb{E}\left[T_q(n)\right] \leq \tau_q(n)$.

Also note that any node in I^c has an ancestor in I and define $\mathcal{I} = \{p \in I^c : p^- \in I\}$ and $\mathcal{I}_h = \{p \in \mathcal{I} : |p| = h\}$.

Thusly motivated we define the partitioning $\mathcal{T} = \mathcal{T}^{(1)} \cup \mathcal{T}^{(2)} \cup \mathcal{T}^{(3)}$ with

$$\begin{aligned}\mathcal{T}^{(1)} &= \{q \succeq p : p \in I_h\}\\ \mathcal{T}^{(2)} &= \bigcup_{h<H} I_h\\ \mathcal{T}^{(3)} &= \bigcup_{h<H} \{q \succeq p : p \in \mathcal{I}_h\}\end{aligned}$$

This decomposition is illustrated in figure 2.2.

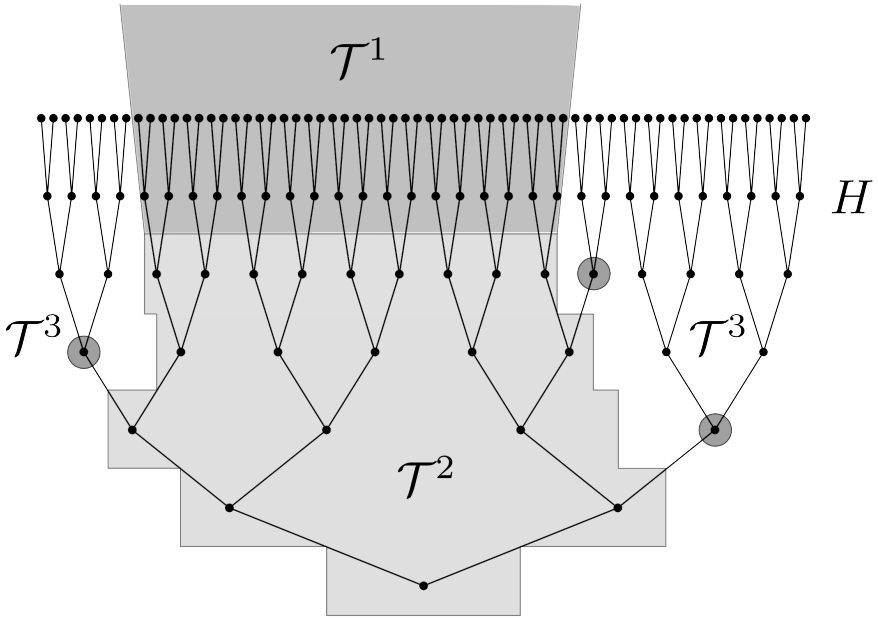


Figure 2.2.: First 7 levels of the decomposition of $\mathcal{T}$ for $H = 6$. Nodes in $\mathcal{I}$ are marked by gray balls.

Now we define the corresponding versions of $\mathcal{T}(n)$ and R_n

$$\mathcal{T}^{(i)}(n) = \mathcal{T}^{(i)} \cap \mathcal{T}(n) \qquad \text{and} \qquad R_n^{(i)} = \sum_{p \in \mathcal{T}^{(i)}(n)} f^* - f\left(X_p\right)$$

and bound $R_n = R^{(1)} + R^{(2)} + R^{(3)}$ by bounding each summand individually.

$\left(\boldsymbol{R^{(1)}}\right)$ For any $p \in \mathcal{T}^{(1)}$ let $p_{\leq H} \in I_H$ be such that $p \succeq p_{\leq H}$. Then

$$\begin{aligned}
\mathbb{E}\left[R_n^{(1)}\right] &= \mathbb{E}\left[\sum_{p\in\mathcal{T}^{(1)}(n)} f^* - f\left(X_p\right)\right] \\
&\leq \mathbb{E}\left[\sum_{p\in\mathcal{T}^{(1)}(n)} \Delta_p\right] \\
&\leq \mathbb{E}\left[\sum_{p\in\mathcal{T}^{(1)}(n)} \Delta_{p_{\leq H}}\right] && \text{lemma 16} \\
&\leq \mathbb{E}\left[\sum_{p\in\mathcal{T}^{(1)}(n)} \delta_H\right] && p_{\leq H} \in I_H \\
&\leq n\delta_H.
\end{aligned}$$

$\left(\boldsymbol{R^{(2)}}\right)$

$$\begin{aligned}
\mathbb{E}\left[R_n^{(2)}\right] &= \mathbb{E}\left[\sum_{p\in\mathcal{T}^{(2)}(n)} f^* - f\left(X_p\right)\right] \\
&\leq \mathbb{E}\left[\sum_{p\in\mathcal{T}^{(2)}(n)} \delta_p\right] && p \in I \\
&\leq \sum_{p\in\mathcal{T}^{(2)}} \delta_p = \sum_{h=1}^{H-1} |I_h|\, \delta_h \leq \sum_{h=1}^{H-1} \eta_h \delta_h
\end{aligned}$$

$\left(\boldsymbol{R^{(3)}}\right)$ First assume, that the bounds τ_p are given and define $\tau_h(n) = \max_{|p|=h} \tau_p(n)$.

Then

$$\begin{aligned}
\mathbb{E}\left[R_n^{(3)}\right] &= \mathbb{E}\left[\sum_{h=1}^{H}\sum_{Q\in\mathscr{I}_h}\sum_{\substack{q\succeq Q\\ q\in\mathscr{T}(n)}} f^* - f\left(X_q\right)\right] \\
&\leq \mathbb{E}\left[\sum_{h=1}^{H}\sum_{Q\in\mathscr{I}_h}\sum_{\substack{q\succeq Q\\ q\in\mathscr{T}(n)}} \delta_{h-1}\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H}\sum_{Q\in\mathscr{I}_h} \delta_{h-1} T_Q(n)\right] \\
&\leq \mathbb{E}\left[\sum_{h=1}^{H} |\mathscr{I}_h|\, \delta_{h-1}\tau_h(n)\right] \\
&\mathscr{I}_h \subseteq \mathscr{C}(I_{h-1}) \Rightarrow |\mathscr{I}_h| \leq k\,|I_{h-1}| \\
&\leq \mathbb{E}\left[\sum_{h=1}^{H} k\,|I_{h-1}|\, \delta_{h-1}\tau_h(n)\right] \\
&\leq k\sum_{h=1}^{H} \eta_{h-1}\delta_{h-1}\tau_h(n)
\end{aligned}$$

The remaining part of the proof will be committed to finding the bounds $\tau_q(n)$. Let $Q \in \mathscr{I}_h$ and assume $Q \preceq q \preceq P_t$. This can only happen if $B_{P(t)_1}(t) \geq B_p(t)$ [13] for all $p \in \mathscr{T}(t)$ with $|p| = 1$ (and in particular for $*[1]$). Therefore

$$\{P_t \succeq Q\} \subseteq \left\{B_{P(t)_1}(t) \geq B_{*[1]}(t)\right\}$$

Recall that by definition

$$p \preceq q \quad \Rightarrow \quad B_p(t) \leq B_q(t) \leq U_q(t)$$

and therefore

$$\begin{aligned}
\left\{B_{P(t)_1}(t) \geq B_{*[1]}(t)\right\} &\subseteq \left\{U_q(t) \geq B_{*[1]}(t)\right\} \\
&\subseteq \left\{U_q(t) \geq B_{*[1]}(t) \wedge \left(B_{*[1]}(t) > f^* \vee B_{*[1]}(t) \leq f^*\right)\right\} \\
&\subseteq \left\{U_q(t) \geq B_{*[1]}(t) > f^*\right\} \cup \left\{U_q(t) \geq B_{*[1]}(t) \wedge B_{*[1]}(t) \leq f^*\right\} \\
&\subseteq \left\{U_q(t) > f^*\right\} \cup \left\{B_{*[1]}(t) \leq f^*\right\}
\end{aligned}$$

[13] Recall that $P_t \in \{1,\dots k\}^{<\omega}$ is a sequence.

Now by the definition of the B-values,

$$\left\{B_{*[h]}(t) \le f^*\right\} \subseteq \left\{U_{*[h]}(t) \le f^*\right\} \cup \left\{B_{*[h+1]}(t) \le f^*\right\}.$$

Since up to round t no more than t nodes have been played, we know that $*[t]$ has not been played so far and thus has a B-value equal to $+\infty$. Therefore we have the inclusion

$$\{P_t \succeq q\} \subseteq \left\{U_q(t) > f^*\right\} \cup \left(\left\{U_{*[1]}(t) \le f^*\right\} \cup \cdots \cup \left\{U_{*[t-1]}(t) \le f^*\right\}\right) \tag{2.6}$$

Similarly as in the proof of theorem 9, we bound

$$\begin{aligned}
\mathbb{E}\left[T_q(n)\right] &\le u + \sum_{t=u+1}^{n} \mathbb{E}\left[P_t \succeq q,\ T_q(t) > u\right] \\
&\le u + \sum_{t=u+1}^{n} \mathbb{P}\left[\left(U_{*[s]}(t) \le f^* \text{ for some } s < t \text{ or } U_q(t) > f^*\right) \text{ and } T_q(t) > u\right] \\
&\le u + \sum_{t=u+1}^{n} \mathbb{P}\left[\left(U_{*[s]}(t) \le f^* \text{ for some } s < t\right) \text{ or } \left(U_q(t) > f^* \text{ and } T_q(t) > u\right)\right] \\
&\le u + \sum_{t=u+1}^{n} \left(\sum_{s=0}^{t-1} \mathbb{P}\left[U_{*[s]}(t) \le f^*\right] + \mathbb{P}\left[U_q(t) > f^* \text{ and } T_q(t) > u\right]\right)
\end{aligned}$$

Now we will bound the probabilities in this sum.

- First we bound $\mathbb{P}\left[U_{*[h]}(n) \le f^*\right]$.

$$\begin{aligned}
&\mathbb{P}\left[U_{*[h]}(n) \le f^*\right] \\
&= \mathbb{P}\left[\hat{\mu}_{*[h]}(n) + \gamma_{*[h]}(n) + \delta_h \le f^*\right] \\
&\quad \text{adding } \mathbb{E}\left[\hat{\mu}_{*[h]}(n) \mid \mathscr{T}(n)\right] \text{on both sides and using that} \\
&\quad \text{by (A}\delta\text{) and remark 15 } \left(\mathbb{E}\left[\hat{\mu}_{*[h]}(n) \mid \mathscr{T}(n)\right] + \delta_h - f^*\right) \ge 0 \\
&\le \mathbb{P}\left[\hat{\mu}_{*[h]}(n) \le \mathbb{E}\left[\hat{\mu}_{*[h]}(n) \mid \mathscr{T}(n)\right] - \gamma_{*[h]}(n)\right] \\
&\le n^{-4} \text{ by (A}\gamma\text{)}
\end{aligned}$$

- Now we bound $\mathbb{P}\left[U_q(t) > f^* \text{ and } T_q(t) > u\right]$.

$$
\begin{aligned}
&\mathbb{P}\left[U_q(t) > f^* \text{ and } T_q(t) > u\right] \\
&= \mathbb{P}\left[\hat{\mu}_q(t) + \gamma_q(t) + \delta_q > f_q^* + \nabla_q \text{ and } T_q(t) > u\right] \\
&\quad f_q^* \geq \frac{1}{T_q(t)} \sum_{s=1}^{t} f\left(X_{P(s)}\right) \{P_s \succeq Q\} = \mathbb{E}\left[\hat{\mu}_q(t) \mid \mathscr{T}(n)\right] \\
&\leq \mathbb{P}\left[\hat{\mu}_q(t) + \gamma_q(t) + \delta_q > \mathbb{E}\left[\hat{\mu}_q(t) \mid \mathscr{T}(n)\right] + \nabla_q \text{ and } T_q(t) > u\right] \\
&= \mathbb{P}\left[\hat{\mu}_q(t) > \mathbb{E}\left[\hat{\mu}_q(t) \mid \mathscr{T}(n)\right] + \left(\nabla_q - \delta_q\right) - \gamma_q(t) \text{ and } T_q(t) > u\right] \\
&\quad \text{union bound on } t \geq T_q(t) > u \\
&\leq \sum_{s=u+1}^{t} \mathbb{P}\left[\hat{\mu}_q(t) > \mathbb{E}\left[\hat{\mu}_q(t) \mid \mathscr{T}(n)\right] + \left(\nabla_q - \delta_q\right) - \gamma_q(t) \text{ and } T_q(t) = s\right]
\end{aligned}
$$

Where $\nabla_q = f^* - f_q^*$.
Since $q \notin I \Rightarrow \left(\nabla_q - \delta_q\right) > 0$, we can choose u such that

$$\gamma_q\left(t, T_q(t)\right) \leq \gamma_q(t,u) \leq \frac{\left(\nabla_q - \delta_q\right)}{2}.$$

Then $\left(\nabla_q - \delta_q\right) - \gamma_q(t) \geq \gamma_q(t)$ and

$$
\begin{aligned}
&\mathbb{P}\left[U_q(t) > f^* \text{ and } T_q(t) > u\right] \\
&\leq \sum_{s=u+1}^{t} \mathbb{P}\left[\hat{\mu}_q(t) \geq \mathbb{E}\left[\hat{\mu}_q(t) \mid \mathscr{T}(n)\right] + \gamma_q(t)\right] \\
&\leq t n^{-4}
\end{aligned}
$$

The condition $\frac{\Delta_q-\delta_q}{2} \geq \gamma_q(t,u) \geq \sqrt{\frac{8\ln n}{u}}$ implies $u \geq \frac{32\ln n}{(\Delta_q-\delta_q)^2}$. So all in all

$$\begin{aligned}
\mathbb{E}\left[T_q(n)\right] &\leq u + \sum_{t=u+1}^{n}\left(\sum_{s=0}^{t-1}\mathbb{P}\left[U_{*[s]}(t) \leq f^*\right] + \mathbb{P}\left[U_q(t) > f^* \text{ and } T_q(t) > u\right]\right) \\
&\leq \frac{32\ln n}{(\Delta_q-\delta_q)^2} + \sum_{t=u+1}^{n}\left(\sum_{s=0}^{t-1} n^{-4} + tn^{-4}\right) \\
&= \frac{32\ln n}{(\Delta_q-\delta_q)^2} + 2\sum_{t=u+1}^{n} tn^{-4} \\
&\leq \frac{32\ln n}{(\Delta_q-\delta_q)^2} + 2\sum_{t=1}^{\infty} t^{-3} \\
&\leq \frac{32\ln n}{(\Delta_q-\delta_q)^2} + 4
\end{aligned}$$

■

Remark 26. UCB1 is a special case of HOO in the sense, that if HOO is initialized with a k-armed bandit $((X_1,\ldots,X_k),M)$ and the parameters

- $\mathcal{X}_p = \{X_{p(1)}\}$ for $|p| > 0$
- $\delta_h = 0$
- $\eta_h = K$
- γ as in corollary 10.

then both algorithms produce the same sequence of pulls.

Remark 27. DOO2 is a special case of HOO where $\gamma = 0$.

3. Specialization to pre-metric spaces

The assumptions (Aδ) and (Dη) are rather technical. In this section we will assume that $\mathcal{X}$ is a pre-metric space and that the tree $\mathcal{T}$ is binary. We will then present assumptions that will simplify the initialization of the algorithms.

A pre-metric space is a set $\mathcal{X}$ together with a function ℓ satisfying

$$(\mathbf{A}\ell) \qquad \ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \text{ such that } \ell(x,x) = 0$$

Example 28. *The Kronecker delta is a prime example of a pre-metric.*
Let $(\mathcal{X}, M)$ be a k-armed bandit, then $(\mathcal{X}, \delta)$ is a pre-metric space.

Given ℓ, we can specify the required continuity to be a weak form of Lipschitz continuity in X^*

$$\left(\mathbf{A}\delta^{\ell}\right) \qquad f^* - f(X) \leq \ell(X, X^*)$$

This condition is illustrated in figure 3.1.
We furthermore assume

$$(\mathbf{A}\rho) \qquad \exists \nu, \rho : \operatorname{diam} \mathcal{X}_p \leq \nu \rho^{|p|}$$

which ensures, that the diameters[1] of optimal cells decrease with depth.

Example 29. *Let the partitioning of $\mathcal{X} = [-1,1]$ be defined as in example 14. The euclidean metric* d *is a pre-metric and satisfies $(A\rho)$ with $\nu = 2$ and $\rho = 2^{-1}$.*

[1] We define $\operatorname{diam} \mathcal{X}_p = \sup \ell\left[\mathcal{X}_p^2\right]$.

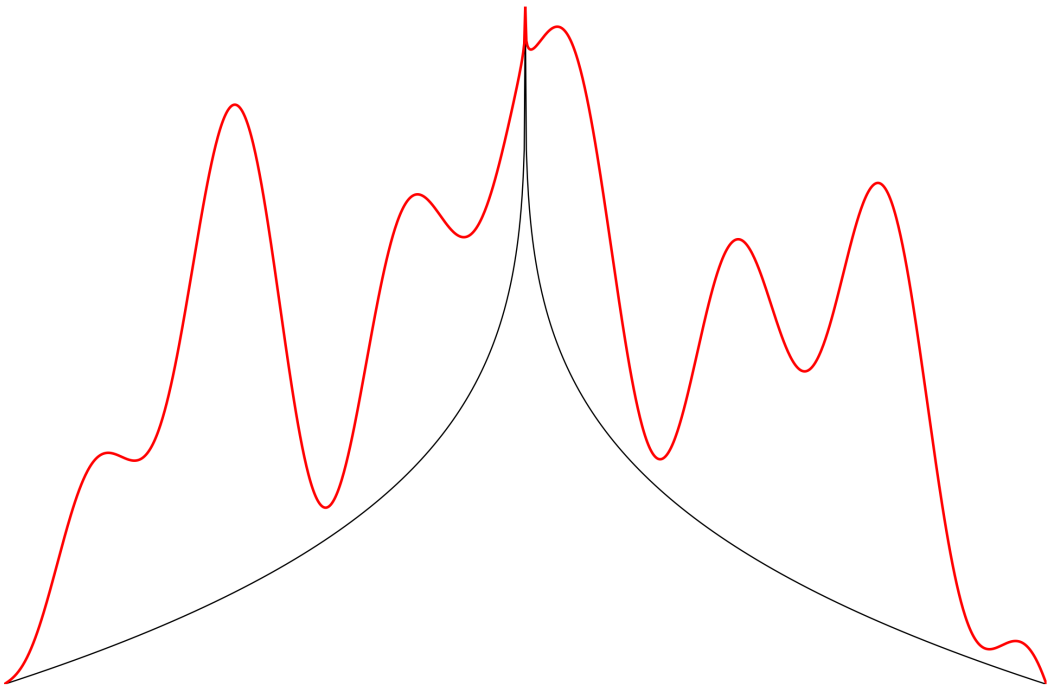


Figure 3.1.: By $\left(\mathrm{A}\delta^{\ell}\right)$, $f(X)$ (red) has to lie above $f^* - \ell(X, X^*)$ (black).

If $\left(\mathrm{A}\delta^{\ell}\right)$ and $(\mathrm{A}\rho)$ are satisfied, then $(\mathrm{A}\delta)$ is satisfied as well. Since for $X, X^* \in \mathscr{X}_{*[h]}$, we can bound

$$f^* - f(X) \leq \ell(X, X^*) \leq \operatorname{diam} \mathscr{X}_{*[h]} \leq \nu\rho^h$$

and define $\delta_p = \nu\rho^{|p|}$. By definition of I_h, this also implies that for all $p \in I_h$

$$\mathscr{X}_p \subseteq \mathscr{X}_{\delta_p} = \mathscr{X}_{\nu\rho^h} \tag{3.1}$$

where $\mathscr{X}_\epsilon := \{X \in \mathscr{X} : f^* - f(X) \leq \epsilon\}$.

Example 30. *Let $\mathscr{X}, f$ and α be as in example 22. Then*

$$f^* - f(X) = |X|^\alpha$$

And if we choose the pre-metric $\ell(X, Y) = |X - Y|^\beta$, then $\left(A\delta^{\ell}\right) \iff |X|^\alpha \leq |X|^\beta$ is

satisfied for any $\beta \leq \alpha$. Finally compute

$$\begin{aligned}\operatorname{diam}\left(\mathscr{X}_p\right) &= \sup\left\{\ell\left(X,Y\right) : X,Y \in \mathscr{X}_p\right\}\\ &= \sup\left\{|X-Y|^\beta : X,Y \in \mathscr{X}_p\right\}\\ &= \left(\sup\left\{|X-Y| : X,Y \in \mathscr{X}_p\right\}\right)^\beta\\ &= \left(2\cdot 2^{-|p|}\right)^\beta \textit{ see example 29}\end{aligned}$$

Therefore $\nu = 2^\beta$ and $\rho = 2^{-\beta}$.

Example 31 (Counterexample)**.** *Even though*

$$\ell(x,y) = \begin{cases} 0 & x = y \\ \frac{1}{\|x-y\|} & x \neq y \end{cases}$$

is a pre-metric, it does not satisfy $(A\rho)$, since $\operatorname{diam}\mathscr{X}_{si} > \operatorname{diam}\mathscr{X}_s$.

Similar to [Munos, 2011, Bubeck et al., 2011] we define an **ℓ-ball** with center x and radius ϵ to be the set $\mathscr{B}_\epsilon(x) = \{y \in \mathscr{X} : \ell(y,x) < \epsilon\}$. To find the bound $(D\eta)$, we assume that

(A$\mathscr{B}$) There exists $\widetilde{\nu} > 0$ and a tree $\mathscr{B}_p \subseteq \mathscr{X}_p$ of ℓ-balls such that

- $\operatorname{diam}\mathscr{B}_p \geq \widetilde{\nu}\rho^{|p|}$ and
- $\mathscr{B}_p \cap \mathscr{B}_q = \emptyset$ for all $p \neq q$ with $|p| = |q|$.

This allows us to bound

$$|I_h| \leq \mathscr{N}^{\ell,\text{pack}}_{\widetilde{\nu}\rho^h}\left(\bigcup_{p\in I_h}\mathscr{X}_p\right) \underset{(3.1)}{\leq} \mathscr{N}^{\ell,\text{pack}}_{\widetilde{\nu}\rho^h}\left(\mathscr{X}_{\nu\rho^h}\right)$$

Where $\mathscr{N}^{\text{pack}}_\epsilon(X)$ denotes the **packing number** of X. It is defined as the maximal number of disjoint balls of radius ϵ and center in X.

Example 32. *Define $\mathscr{X}$, ℓ and β as in example 30.*
Then $\mathscr{X}_p$ are disjoint ℓ-balls of radius $2^\beta \cdot 2^{-\beta|p|}$. So $\widetilde{\nu} = \nu = 2^\beta$.

The remainder of this section is devoted to the search for a bound for $\mathscr{N}^{\ell,\text{pack}}_{\widetilde{\nu}\rho^h}\left(\mathscr{X}_{\nu\rho^h}\right)$. For this we will need the **covering number** $\mathscr{N}^{\ell,\text{cover}}_\epsilon(X)$. It is defined as the minimal number of ℓ-balls of radius ϵ and center in X such that their union contains X. If a statement is true for the packing, as well as the covering number, then we will omit the second superscript.

Lemma 33. *Let $\epsilon, \zeta \geq 0$ and $X, X_1, X_2 \subseteq \mathscr{X}$, then $(A\rho)$ implies*

$$\mathscr{N}_\epsilon^\ell(X) \geq \mathscr{N}_{\epsilon+\zeta}^\ell(X)$$

$$\mathscr{N}_\epsilon^\ell(X_1 \cup X_2) \leq \mathscr{N}_\epsilon^\ell(X_1) + \mathscr{N}_\epsilon^\ell(X_2)$$

And, if ℓ is symmetric

$$\mathscr{N}_\epsilon^{\ell,pack}(X) \leq \mathscr{N}_\epsilon^{\ell,cover}(X)$$

Proof.

- We will prove this for the covering number. Let $\{x_i\}$ be the centers of an ϵ-covering of X. Since $B_\epsilon(x_i) \subseteq B_{\epsilon+\zeta}(x_i)$, these are also centers for an $(\epsilon+\zeta)$-covering.
- We will prove this for the covering number. Let $\{B_{i,j}\}$ be a covering of X_j. Then $\{B_{i,1}\} \cup \{B_{i,2}\}$ is a covering for $X_1 \cup X_2$.
- Let $\{x_i\}$ and $\{y_i\}$ be the centers of the maximal ϵ-packing and the minimal ϵ-covering of X respectively. Then $\forall x_i \exists y_j : x_i \in B_\epsilon(y_j)$.
  Since ℓ is assumed to be symmetric, $x_i \in B_\epsilon(y_j)$ implies $y_j \in B_\epsilon(x_i)$.
  So for each x_i the set $\{y_j : y_j \in B_\epsilon(x_i)\}$ is non-empty. But since the $B_\epsilon(x_i)$ are disjoint so must be these sets. Therefore $|\{x_i\}| \leq |\{y_j\}|$.

■

In the following we will assume ℓ to be symmetric.[2]

Define the **doubling constant** to be the minimal number $q \geq 1$ such that every ball in $\mathscr{X}$ can be covered q balls of half its radius.
As is the case in euclidean spaces, we assume it to be finite.[3] A simple calculation shows, that for any ℓ-ball X and any $k \geq 0$

$$\mathscr{N}_{2^{-k}\epsilon}^\ell(X) \leq q^{\lceil k \rceil} \mathscr{N}_\epsilon^\ell(X) \tag{3.2}$$

Example 34. *Let $(\mathscr{X}, \mathrm{d})$ be the D-dimensional euclidean space. Then it's doubling constant is 2^D.*

Example 35. *Define $\mathscr{X}, \ell$ and β as in example 30 and let d be the euclidean metric on $\mathbb{R}$. Then*

$$\ell(X,Y) \leq \epsilon \Leftrightarrow \mathrm{d}(X,Y) \leq \epsilon^{1/\beta}$$

[2] If ℓ is a pre-metric which satisfies $(\mathrm{A}\rho)$ and $(\mathrm{A}\delta^\ell)$ with constants ν and ρ, then so is $\ell^{\mathrm{sym}}(x,y) := \ell(x,y) \vee \ell(y,x)$.

[3] This does not have to be the case in general.

and consequently $\mathcal{N}_\epsilon^\ell(X) = \mathcal{N}_{\epsilon^{1/\beta}}^{\mathrm{d}}(X)$ *for any ball* $X \subseteq \mathcal{X}$.

Now define $c = 2^{-1/\beta}$. *Then*

$$\begin{aligned}
\mathcal{N}_{\epsilon/2}^\ell(X) &= \mathcal{N}_{c\epsilon^{1/\beta}}^{\mathrm{d}}(X) \\
&\le \mathcal{N}_{2^{\lfloor \mathrm{ld}\, c\rfloor}\epsilon^{1/\beta}}^{\mathrm{d}}(X) && \textit{by lemma 33} \\
&\le 2^{-\lfloor \mathrm{ld}\, c\rfloor}\mathcal{N}_{\epsilon^{1/\beta}}^{\mathrm{d}}(X) && \textit{see example 34} \\
&= 2^{-\lfloor \mathrm{ld}\, c\rfloor}\mathcal{N}_\epsilon^\ell(X).
\end{aligned}$$

Therefore the doubling constant of ℓ *is* $q = 2^{-\lfloor -1/\beta\rfloor} = 2^{\lceil 1/\beta\rceil}$.

Lemma 36. $\mathcal{N}_{\epsilon/2}^{\ell,cover}(X) \le q\mathcal{N}_\epsilon^{\ell,cover}(X)$ *for any* $X \subseteq \mathcal{X}$.

Proof. Let B_i be a minimal ϵ-covering for X.

$$\begin{aligned}
\mathcal{N}_{\epsilon/2}^{\ell,\mathrm{cover}}(X) &\le \sum_{i=1}^{\mathcal{N}_\epsilon^{\ell,\mathrm{cover}}(X)} \mathcal{N}_{\epsilon/2}^{\ell,\mathrm{cover}}(B_i) \le \sum_{i=1}^{\mathcal{N}_\epsilon^{\ell,\mathrm{cover}}(X)} q\mathcal{N}_\epsilon^{\ell,\mathrm{cover}}(B_i) \\
&= q \sum_{i=1}^{\mathcal{N}_\epsilon^{\ell,\mathrm{cover}}(X)} 1 = q\mathcal{N}_\epsilon^{\ell,\mathrm{cover}}(X)
\end{aligned}$$

■

Theorem 37. *Assume that* $\mathcal{N}_{\tilde{\nu}}^{\ell,cover}(\mathcal{X}_\nu)$ *is finite.*
Then there exist d *such that* $|I_h| \le C\rho^{-dh}$.

The smallest d *with this property is called the* ***near-optimality dimension****.*

Proof. First observe that by assumption and lemma 33

$$|I_h| \le \mathcal{N}_{\tilde{\nu}\rho^h}^{\ell,\mathrm{pack}}\left(\mathcal{X}_{\nu\rho^h}\right) \le \mathcal{N}_{\tilde{\nu}\rho^h}^{\ell,\mathrm{cover}}\left(\mathcal{X}_{\nu\rho^h}\right).$$

Now by induction on h show that $\mathcal{N}_{\tilde{\nu}\rho^h}^{\ell,\mathrm{cover}}\left(\mathcal{X}_{\nu\rho^h}\right) \le C\rho^{-dh}$.

$\boxed{h = 0}$ Define $C := \mathcal{N}_{\tilde{\nu}}^{\ell,\mathrm{cover}}(\mathcal{X}) \ge \mathcal{N}_{\tilde{\nu}}^{\ell,\mathrm{cover}}(\mathcal{X}_\nu)$ which is finite by assumption.

$\boxed{h \rightsquigarrow h+1}$ Let $u \in \mathbb{N}$, $d \in \mathbb{R}$ be such that $2^{-u} \le \rho$ and $\rho^{-d} = q^u$. Then

$$\begin{aligned}
\mathcal{N}_{\tilde{\nu}\rho^{h+1}}^{\ell,\mathrm{cover}}\left(\mathcal{X}_{\nu\rho^{h+1}}\right) &\le \mathcal{N}_{2^{-u}\tilde{\nu}\rho^h}^{\ell,\mathrm{cover}}\left(\mathcal{X}_{\nu\rho^h}\right) && \text{by lemma 33} \\
&\le q^u\mathcal{N}_{\tilde{\nu}\rho^h}^{\ell,\mathrm{cover}}\left(\mathcal{X}_{\nu\rho^h}\right) && \text{by lemma 36} \\
&\le q^u C\rho^{-dh} && \text{IH} \\
&= C\rho^{-d(h+1)}
\end{aligned}$$

■

Remark 38. *For this bound, it suffices that there exists some h for which* $\mathcal{N}^{\ell,cover}_{\tilde{\nu}\rho^h}\left(\mathcal{X}_{\nu\rho^h}\right)$ *is finite. This h does not have to be* 0.

3.1. Regret bound for DOO

Theorem 39 (Regret bound vor DOO)**.**

$\boxed{d=0}$ $R_n \in \mathcal{O}(1-\rho^n)$

$\boxed{d\in(0,1)}$ $R_n \in \mathcal{O}\left(1-n^{(1-d)\ln\rho}\right)$

$\boxed{d=1}$ $R_n \in \mathcal{O}(\ln n)$

$\boxed{d>1}$ $R_n \in \mathcal{O}\left(n^{(1-d)\ln\rho}\right)$

Proof. From theorem 20 we know, that

$$r_n \leq \delta_{h(n)} \leq \nu\rho^{h(n)}$$
$$R_n \leq \sum_{l=0}^{h(n)} \eta_l \delta_l \leq C\nu \sum_{l=0}^{h(n)} \rho^{(1-d)l}$$

where $h(n) = \inf\left\{h : \sum_{l=0}^{h} C\rho^{-dl} \geq n\right\}$.
If $\rho^{-d}=1$ then $h(n) = \left\lceil \frac{n}{C} \right\rceil - 1$. Otherwise we can use the well-known identity

$$\sum_{l=0}^{h} C\rho^{-dl} = C\frac{\rho^{-dh}-1}{\rho^{-d}-1}$$

for partial sums of geometric series to conclude $h(n) = \ln\left(n\left(\frac{\rho^{-d}-1}{C}\right)+1\right)/\ln\left(\rho^{-d}\right)$.

This gives:

$\boxed{d=0}$ $h(n) \in \mathcal{O}(n)$

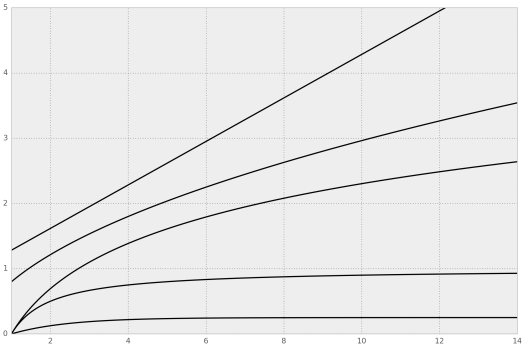
$\boxed{d\neq 0}$ $h(n) \in \mathcal{O}(\ln n)$

Finally we can bound for $R_n \in \mathcal{O}(\ln n)$ for $d=1$ and otherwise

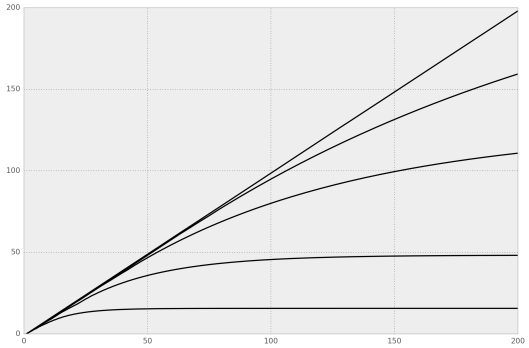
$$R_n \leq C\nu\frac{\rho^{(1-d)h(n)}-1}{\rho^{(1-d)}-1}$$

This concludes the proof.

■



Theoretical bound for $d \in \left\{o, \frac{1}{2}, 1, \frac{3}{2}, 2\right\}$ (from bottom to top).



Experimental values for $\mathcal{X}, f$ and different β defined as in example 22.

Figure 3.2.: Theoretical (left) and experimental bounds for R_n for DOO .

3.2. Regret bound for HOO

Theorem 40 (Regret bound for HOO)**.**

$$\mathbb{R}[R_n] \in \mathcal{O}\left(n^{(d+1)/(d+2)} (\ln n)^{1/(d+2)}\right)$$

Proof. From theorem 25 we know that

$$\mathbb{E}[R_n] \leq n\delta_H + \sum_{h=0}^{H-1} \eta_h \delta_h + k \sum_{h=1}^{H} \eta_{h-1} \delta_{h-1} \tau_h$$

where $\tau_h = \mathcal{O}(\ln n)$. Therefore

$$\begin{aligned}
\mathbb{E}[R_n] &\leq \mathcal{O}\left(n\rho^H\right) + \mathcal{O}\left(\sum_{h=0}^{H-1} \rho^{h(1-d)}\right) + \mathcal{O}\left((\ln n) \sum_{h=1}^{H} \rho^{-h(d+1)}\right) \\
&\leq \mathcal{O}\left(n\rho^H\right) + \mathcal{O}\left(\rho^{-(H+1)(d+1)}\right) + \mathcal{O}\left((\ln n)\, \rho^{-(H+1)(d+1)}\right) \\
&\leq \mathcal{O}\left(n\rho^H + (\ln n)\, \rho^{-H(d+1)}\right)
\end{aligned}$$

Now by defining $x = \rho^H$ and $e = d + 1$ we can rewrite

$$n\rho^H + (\ln n)\, \rho^{-H(d+1)} = nx + x^{-e} \ln n.$$

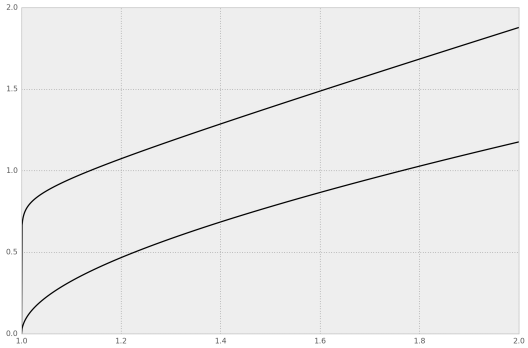
Since $H \in \mathbb{R}$ was arbitrary we can choose it such that x takes any value in $\mathbb{R}_+$. So for a fixed n we can choose $x = n^a (\mathrm{ld}\, n)^b$ and obtain the bound

$$\mathbb{E}[R_n] \leq \mathcal{O}\left(n^{1+a} (\ln n)^b + n^{-ae} (\mathrm{ld}\, n)^{1-be}\right).$$

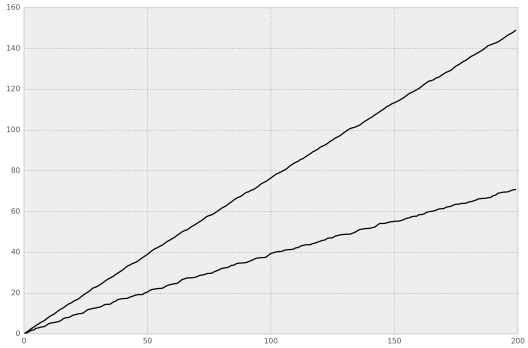
By choosing a and b such that $1 + a = -ae$ and $b = 1 - be$ we obtain

$$\mathbb{E}[R_n] \leq \mathcal{O}\left(n^{(d+1)/(d+2)} (\ln n)^{1/(d+2)}\right)$$

■



Theoretical bound for $d = 0$ (bottom) and $d = 15$ (top).



Experimental values for $\mathcal{X}, f$ and different β defined as in example 22.

Figure 3.3.: Theoretical (left) and experimental bounds for R_n for HOO .

Example 41. *Let $\mathcal{X} = [0,1]^D$ and define the hierarchical partitioning by recursively dividing each hypercube $\mathcal{X}_p$ into 2^D hypercube with sides of half the length of $\mathcal{X}_p$ and X_p to be the center of $\mathcal{X}_p$. Define*

$$f(X) = 1 - \|X\|^\alpha \qquad \text{and} \qquad \ell(X,Y) = \|X\|_\infty^\beta$$

where again $\beta \leq \alpha$. It is an easy exercise to compute

$$\nu = \tilde{\nu} = 2^\beta \qquad \rho = 2^{-\beta} \qquad q = 2^{D\lceil 1/\beta \rceil}$$

Guided by the proof of theorem 37 we now compute a first bound on the near-optimality dimension. We have to find $u \in \mathbb{N}$ and $d' \in \mathbb{R}$ such that $2^{-u} \leq \rho$ and $\rho^{-d} = q^u$:

$$\begin{array}{rrcl@{\qquad}rcl} & 2^{-u} & \leq & \rho & \rho^{-d'} & = & q^u \\ \Leftrightarrow & -u & \leq & \mathrm{ld}\, \rho & -d' \mathrm{ld}\, \rho & = & u \,\mathrm{ld}\, q \\ \Leftarrow & u & = & \lceil -\mathrm{ld}\, \rho \rceil & d' & = & \frac{u \,\mathrm{ld}\, q}{-\mathrm{ld}\, \rho} \end{array}$$

Therefore $d \leq d' = \dfrac{\lfloor \operatorname{ld} \rho \rfloor \operatorname{ld} q}{\operatorname{ld} \rho} \leq \operatorname{ld} q = D \lceil 1/\beta \rceil$.

This bound however is not tight.
Indeed [Munos, 2011] computed $d = D(1/\beta - 1/\alpha)$ *directly from* $|I_h| \leq \mathcal{N}_{\nu\rho^h}^{\ell,pack}\left(\mathcal{X}_{\nu\rho^h}\right)$.

Let $\epsilon = \nu\rho^h$. *First observe that* $\mathcal{X}_\epsilon$ *is the* $\|\bullet\|_\infty$*-ball of radius* $\epsilon^{1/\alpha}$ *centered in* 0. *Secondly observe that we want to pack this ball by* ℓ*-balls of radius* ϵ *and that those are* $\|\bullet\|_\infty$*-balls of radius* $\epsilon^{1/\beta}$. *Therefore we can achieve a packing of at most* $\left(\frac{\epsilon^{1/\alpha}}{\epsilon^{1/\beta}}\right)^D$ *balls and for* $d = D(1/\beta - 1/\alpha)$ *we can bound*

$$|I_h| \leq \nu^d \rho^{-dh}$$

This demonstrates that the bound we found in the proof of theorem 37 is more of theoretical than of practical interest.

Remark 42. *In theorem 39 and theorem 40 we have seen that the growth of the regret is heavily influenced by* d *and it is advantageous to get* d *as low as possible.*

Then in example 41 we have seen, that achieving $d = 0$ *has the additional advantage that in this case the regret is independent of the dimension of the space.*

Achieving $d = 0$ *however is not always possible. Consider for example the function*

$$f(X) := 1 + s(\operatorname{ld}|X|) \cdot \left(\sqrt{|X|} - x^2\right) - \sqrt{|X|}$$

where, $s(x) = \lceil x \rceil - \operatorname{round}(x)$. *(depicted in Figure 3.4(a))*

This example was first given by [Grill et al., 2015] and [Valko et al., 2013]. To see, that there indeed is no semi-metric of the form $\ell(X,Y) = \|X - Y\|^\beta$ *for which the near-optimality dimension of* f *is* $d = 0$ *first define*

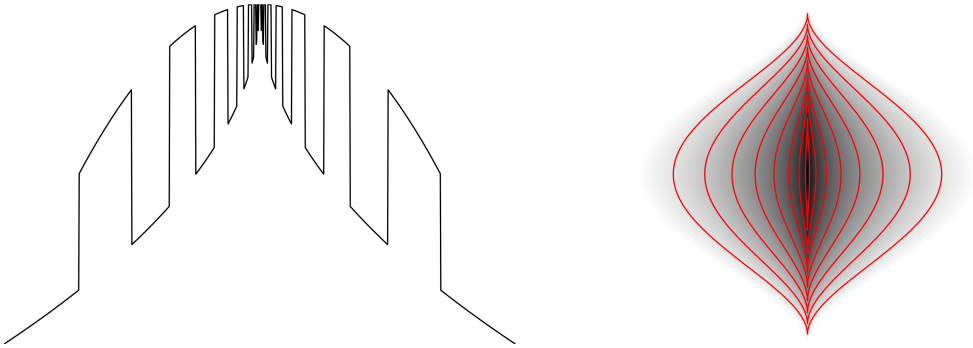
$$\mathcal{X}^{1/2} = \{X \in \mathcal{X} : \operatorname{round}(\operatorname{ld}|X|) \neq \lceil \operatorname{ld}|X| \rceil\} \qquad \text{and} \qquad \mathcal{X}^2 = \mathcal{X} \setminus \mathcal{X}^{1/2}.$$

Then we can write f *as* $f(X) = 1 - \left\{X \in \mathcal{X}^{1/2}\right\}\sqrt{X} - \left\{X \in \mathcal{X}^2\right\}X^2$ *and know from example 41 that for* $d = 0$ *we would need* $\beta = 1/2$ *on* $\mathcal{X}^{1/2}$ *but* $\beta = 2$ *on* $\mathcal{X}^2$.

Other examples of nonzero near-optimality dimension are functions that grow with different rates depending on the direction (see [Bubeck et al., 2011, Grill et al., 2015, Valko et al., 2013]), for instance

$$f(X,Y) := 1 - |X| - Y^2$$

This function is depicted in Figure 3.4(b).



(a) (b)

Figure 3.4.

A. Hoeffding's inequality

Theorem 43 (Azuma-Hoeffding inequality)**.** *Let S_n be a martingale with $S_0 = 0$ and $|S_n - S_{n-1}| \leq \sigma_n$. Then for all $\alpha > 0$*

$$\mathbb{P}[S_n \geq \alpha] \leq \exp\left(-\alpha^2 \Big/ 2\sum_{j=1}^{n} \sigma_j^2\right)$$

Proof. See [Lalley, 2013]. ■

Corollary 44. *Let $X_1, \ldots, X_n \in [0,1]$ be i.i.d. random variables with mean μ and define $\hat{\mu}(n) = \frac{1}{n}\sum_{l=1}^{n} X_l$. Then for all $\alpha \geq 0$, $[\hat{\mu}(n) - \alpha, \hat{\mu}(n) + \alpha]$ is a confidence interval for μ with confidence level $1 - 2\exp\left(-\frac{\alpha^2}{2n}\right)$.*

Proof. Define $S_n = \hat{\mu}(n) - \mu$. It is not hard to see, that S_n and $-S_n$ are martingales. Therefore

$$\begin{aligned}\mathbb{P}[\hat{\mu}(n) - \alpha < \mu < \hat{\mu}(n) + \alpha] &= 1 - \mathbb{P}[S_n - \alpha \geq 0 \text{ or } 0 \geq S_n + \alpha] \\ &\geq 1 - \left(\exp\left(-\frac{\alpha^2}{2n}\right) + \exp\left(-\frac{\alpha^2}{2n}\right)\right)\end{aligned}$$

■

B. Implementation

The complete source code (images and experiments) is written in python and can be provided upon request.

Bibliography

[wik, a] Multi-armed bandit. https://en.wikipedia.org/wiki/Multi-armed_bandit. Retrieved August 4, 2016.

[wik, b] Slot machine. https://en.wikipedia.org/wiki/Slot_machine. Retrieved August 4, 2016.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256.

[Bubeck et al., 2011] Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695.

[Grill et al., 2015] Grill, J.-B., Valko, M., and Munos, R. (2015). Black-box optimization of noisy functions with unknown smoothness. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 667–675. MIT Press.

[Lalley, 2013] Lalley, S. P. (2013). Concentration inequalities. Lecture notes. University of Chicago. https://galton.uchicago.edu/~lalley/Courses/386/Concentration.pdf.

[Munos, 2011] Munos, R. (2011). Optimistic optimization of a deterministic function without the knowledge of its smoothness. In Shawe-taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 783–791.

[Tekin, 2013] Tekin, C. (2013). *Online Learning in Bandit Problems*. Dissertation, University of Michigan.

[Valko et al., 2013] Valko, M., Carpentier, A., and Munos, R. (2013). Stochastic simultaneous optimistic optimization. In Dasgupta, S. and Mcallester, D., editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 19–27. JMLR Workshop and Conference Proceedings.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den 19. Oktober 2016 .....