X -ARMED BANDITS: OPTIMIZING QUANTILES, AND OTHER RISKS

Léonard Torossian, Aurélien Garivier, Victor Picheny

$\mathcal{X}$-ARMED BANDITS: OPTIMIZING QUANTILES, AND OTHER RISKS

Léonard Torossian
MIAT, Université de Toulouse, INRA
and Institut de Mathématiques de Toulouse
leonard.torossian@inra.fr

Aurélien Garivier
Univ. Lyon, ENS de Lyon
aurelien.garivier@ens-lyon.fr

Victor Picheny
PROWLER.io, 72 Hills Road, Cambridge
victor@prowler.io

April 16, 2019

ABSTRACT

We propose and analyze StoROO, an algorithm for risk optimization on stochastic black-box functions derived from StoOO. Motivated by risk-averse decision making fields like agriculture, medicine, biology or finance, we do not focus on the mean payoff but on generic functionals of the return distribution, like for example quantiles. We provide a generic regret analysis of StoROO. Inspired by the bandit literature and black-box mean optimizers, StoROO relies on the possibility to construct confidence intervals for the targeted functional based on random-size samples. We explain in detail how to construct them for quantiles, providing tight bounds based on Kullback-Leibler divergence. The interest of these tight bounds is highlighted by numerical experiments that show a dramatic improvement over standard approaches.

Keywords Optimistic optimization · Risk-averse solutions · Quantile optimization

1 Introduction

We consider an unknown function $\Phi : \mathcal{X} \times \Omega \to [0,1] \subset \mathbb{R}$, where $\mathcal{X} \subset [0,1]^D$ and Ω denotes the probability space representing some uncontrollable variables. For any fixed $x \in \mathcal{X}$, $Y_x = \Phi(x, \cdot)$ is a random variable of law $\mathbb{P}_x$ and we consider $g(x) = \psi(\mathbb{P}_x)$ with ψ, a real-valued functional defined on probability measures. We assume that there exists at least one $x^* \in \mathcal{X}$ such that $g(x^*) = \sup_{x \in \mathcal{X}} g(x)$. Using a set of sequential observations $(g(x_1), \cdots, g(x_T))$, our goal is to minimize the simple regret $r_T = g(x^*) - g(x_T)$, with x_T the value returned after using a budget T.

Different families of algorithms have been developed to treat this problem. Some are for example of Bayesian flavor [see Shahriari et al., 2016, for instance], some are inspired by the bandit literature. Here we focus our interest on the bandit framework.

In the classical $\mathcal{X}$-armed bandit problem, a forecaster selects repeatedly a point x in the input space $\mathcal{X} \in [0,1]^D$ and receives a reward distributed according to an unknown distribution $\mathbb{P}_x$. Historically, the main goal was to minimize the *cumulative regret*, i.e. the sum of the difference between his collected rewards and the ones that would have been brought by optimal actions. In the last decade, other works focused on the simple regret. These can be divided in two: algorithms that optimize an unknown function with the knownledge of the smoothness, for example StoOO [Munos et al., 2014], HOO [Bubeck et al., 2011] Zooming [Kleinberg et al., 2008] or HCT [Azar et al., 2014], and others focusing on the optimization of unknown functions without the knowledge of the smoothness, such as POO [Grill et al., 2015], StroquOOL [Bartlett et al., 2018], GPO [Shang et al., 2019] StoSOO [Valko et al., 2013] or Locatelli and Carpentier [2018].

Those algorithms focus on the optimization of the conditional *expectation* of $\mathbb{P}_x$. This choice is questionable in some situations. For example if the shape and variance of the reward distribution depend on the input, a forecaster may be interested in different aspects of the unknown distribution in order to modulate its risk exposure. In the literature, some measures of risk have been proposed to replace the expectation: for instance quantiles [also referred to as Value-at-Risk, see Artzner et al., 1999, McNeil and Frey, 2000] for instance), the Conditional Value-at-Risk [CVaR, Rockafellar et al., 2000], the entropy Value-at-Risk [Ahmadi-Javid, 2012], or expectiles [Bellini and Di Bernardino, 2017]. The purpose of this paper is to present a risk optimization framework of an unknown stochastic function with the knowledge of the smoothness using only pointwise sequential observations and a finite budget T.

$\mathcal{X}$-armed bandit algorithms rely on *optimistic strategies* that associate with each point of the space an upper confidence bound (UCB), that is, an "optimistic" prediction of the outcome. Adapting the classical setting to the optimization of risk measures implies being able to create high-probability confidence bounds for that particular measure. This problem has been tackled in the multi-armed bandit setting ($i.e.$ when the input space is discrete and finite). For instance, Audibert et al. [2009], Sani et al. [2012] focused on the empirical variance, Galichet et al. [2013], Kolla et al. [2019], Hepworth [2017] on the CVaR while in David and Shimkin [2016], Szorenyi et al. [2015] the authors based their policies on the quantile. However, the literature is scarce in the continuous input space case.

In this paper we provide a new version of the Stochastic Optimistic Optimization (StoOO) algorithm [Munos et al., 2014], named StoROO (Stochastic Risk Optimistic Optimization), which is designed to optimize any function $g(x) = \psi(\mathbb{P}_x)$. In a first part, we provide an analysis of the simple regret from a generic point of view (that is, for any ψ). Then, we apply StoROO to optimize the conditional quantile. Using only the assumption that the output distribution support is connected and bounded in $[0, 1]$ and admits a continuous density, we first propose an upper bound on the simple regret using Hoeffding's inequality. Next we derive confidence intervals that take into account the order of the quantile respectively based on Bernstein's and Chernoff's inequalities. Finally, we present numerical experiments that illustrate the ability of our method to optimize conditional quantiles of a black-box function and the relevance to use confidence bounds derived from Chernoff's inequality. Due to space limitation, technical proofs are deferred to Supplementary Material.

2 Problem setup

2.1 Hierarchical partitioning

The upper confidence bounds on which optimistic algorithms are based are surrogate functions $U : \mathcal{X} \to \mathbb{R}$ larger than the objective (in a sense detailed below) with high probability. At each round t, the point $X(t)$ having the highest UCB is sampled and a reward $Y_X(t)$ is collected.

In the classical multi-armed bandit problem, computing and sorting the UCB can be done without major issues. But dealing with continuous input spaces ($i.e.$ infinitely many arms) implies maximizing a UCB function over a continuous space, which can be both computational intensive and algorithmically challenging. For example, Piyavskii's algorithm [see Bouttier, 2017, and references therein] defines U using a global Lipschitz assumption on the targeted function. Because of the Lipschitz hypothesis, the UCB maximizer is at an intersection of hyperplanes, i.e. where the UCB is non-differentiable. Thus a gradient-based algorithm cannot be used, implying that finding the point with the highest UCB is a very hard problem to solve.

To overcome the computational difficulties, a popular alternative is to rely on hierarchical partitions [Bubeck et al., 2011, Munos et al., 2014]. Let us consider an infinite hierarchical space structure $\mathcal{P} = \{\mathcal{P}_{h,j}\}_{h,j}$ of $\mathcal{X}$ such that

$$\mathcal{P}_{0,1} = \mathcal{X}, \;\; \mathcal{P}_{h,j} = \bigcup_{i=0}^{K-1} \mathcal{P}_{h+1,Kj-i},$$

with K the number of sub-regions obtained after expanding a cell and $\mathcal{P}_{h,j}$ the j-th cell at depth h. In the following we assume that:

Assumption 1: There exists a decreasing sequence $\delta(h)$, such that for any $h \geq 0$ and for any cell $\mathcal{P}_{h,j}$, $\sup_{x \in \mathcal{P}_{h,j}} \|x - x_{h,j}\|_\infty \leq \delta(h)$, with $x_{h,j}$ the center of $\mathcal{P}_{h,j}$.

Assumption 2: There exists $\nu > 0$ such that all cells of depth h contain a ball of radius $\nu\delta(h)$.

Starting with $\mathcal{P}_{0,1}$ and following an optimistic strategy, at time t the algorithm has expanded some cells and the result is a tree $\mathcal{T}_t$ that is a subset of $\mathcal{P}$ and a partition of $\mathcal{X}$. In this setting U is taken as a piecewise constant function. Indeed for any $(\mathcal{P}_{h,j})_{h,j \in \mathcal{T}_t}$ we define $\bar{U}_{h,j}$ such that for all $x \in \mathcal{P}_{h,j}$, $U(x) = \bar{U}_{h,j}$.

In the literature of $\mathcal{X}$-armed bandits there are two ways to select a cell of $\mathcal{T}_t$ at each round. In Bubeck et al. [2011], the algorithm follows an *optimistic path* from the root to the leaves. In Munos et al. [2014], StoOO selects the cell having the highest UCB among all the cells of $\mathcal{T}_t$ that have not been expanded, $i.e.$ the set $\mathcal{L}_t$ of leaves of $\mathcal{T}_t$. We consider here this second alternative. Hence, to find the maximizer of U at time t, we only need to evaluate and sort a finite number of values $(\bar{U}_{h,j})_{(h,j)\in\mathcal{L}_t}$.

2.2 Upper and lower confidence bounds, bias

To create confidence bounds for $(\mathcal{P}_{h,j})_{(h,j)\in\mathcal{L}_t}$, the idea of StoOO is to get a sample of every node cell center $x_{h,j}$. Thanks to the fact that all observed values are independent, we can use a *deviation inequality* to create a UCB for $g(x_{h,j})$, that we denote $U_{h,j}$. Finally to create the UCB over the cell $\bar{U}_{h,j}$, a *bias term* is added that takes into account how g can potentially increase from the center of the cell to its edges.

To ensure the convergence of StoOO (and StoROO), the function $\bar{U}_{h,j}$ only needs to be a UCB of $\max_{x\in\mathcal{P}_{h,j}} g(x)$ for the cell containing x^*, as is detailed in the proof of Proposition 1 (see also Munos et al. [2014]). Bounding by how much g can potentially increase from the center to the edge of the optimal cell requires a regularity assumption on g. Following Munos et al. [2014], Azar et al. [2014], we assume the following smoothness property:

$$\forall x \in \mathcal{X}, \quad g(x) \geq g(x^*) - \beta||x - x^*||^\gamma \text{ with } \gamma,\ \beta\ > 0. \tag{1}$$

Note that this condition is less restrictive than a global Lipschitz condition. It does not exclude functions that are very irregular (possibly discontinuous), except close to global maxima. Based on (1) we define

$$\bar{U}_{h,j} = U_{h,j} + B_{h,j}, \text{ with } B_{h,j} = \beta\delta(h)^\gamma.$$

The algorithm also needs a quantity that bounds g from below in order to provide guaranties on the value of g over each cell. We thus construct a lower confidence bound, termed $L_{h,j}$, for $g(x_{h,j})$, and use it as an LCB for the maximum of g on $\mathcal{P}_{h,j}$. In particular, on the cell $\mathcal{P}_{h^*,j^*}$ containing the optimum x^*, it holds that

$$L_{h^*,j^*} \leq g(x^*) \leq U_{h^*,j^*} + \beta\delta(h^*)^\gamma$$

with high probability. To summarize, the estimation of $g(x^*)$ is altered by two sources of error: the local estimation error $E_{h^*,j^*} = U_{h^*,j^*} - L_{h^*,j^*}$ made at the center of the cell, and the bias term B_{h^*,j^*}. Balancing those two terms naturally provides a trade-off between exploration and exploitation.

3 Stochastic Risk Optimistic Optimization

3.1 The StoROO algorithm

StoROO starts by sampling one time each K sub-region of the root node. Then, at each time $1 \leq t \leq T$ the algorithm selects $\mathcal{P}_{h_t,j_t} \in (\mathcal{P}_{h,j})_{(h,j)\in\mathcal{L}_t}$ having the highest UCB. To reduce the estimation error, StoROO can either get more samples from $\mathcal{P}_{h_t,j_t}$ (to reduce the variance), or split the cell in order to reduce its diameter (to reduce the bias). The good balance between these two options is found by dividing a cell as soon as the local estimation error is smaller than the bias, that is when

$$U_{h_t,j_t} - L_{h_t,j_t} \leq \beta\delta(h_t)^\gamma. \tag{2}$$

If Condition (2) is satisfied, StoROO expands $\mathcal{P}_{h_t,j_t}$ and requires a new sample at the center of each sub-region. If Condition (2) is not satisfied, then StoROO requires a new sample at the center x_{h_t,j_t} which is used to update U_{h_t,j_t} and L_{h_t,j_t}.

When the budget is exhausted, several choices are possible for the return value: they have the same theoretical guarantees. Following Munos et al. [2014], one can return the deepest node among those that have been expanded. Here we propose a different, more conservative choice. Denoting by $\mathscr{L}_T$ the set of nodes having the highest LCB among those that have been expanded after a budget T, StoROO returns the node with the highest value $\hat{g}$ (an estimator of g) among the deepest nodes of $\mathscr{L}_T$.

The pseudo-code of the full algorithm is given in Algorithm 1. It requires the parameters β and γ of Condition (1), but of course the inequality do not have to be tight.

3.2 Analysis of the algorithm

In this section we provide a theoretical analysis of StoROO. It is inspired by Munos et al. [2014], but differs most notably by the fact that the analysis is suited for any g and not only for the conditional expectation.

Algorithm 1 StoROO

Input: error probability $\eta > 0$; number of children K; time horizon T; $\beta > 0$; $\gamma > 0$;
Define: UCB and LCB
Initialization $n = 1$; $t = 1$;
Expand into K sub-regions the root node $(0, 0)$ and sample one time each child
while $n \leq T$ **do**
  foreach $(h, j) \in \mathcal{L}_t$ **do**
    compute $\bar{U}_{h,j}(t)$
  end
  Select $(\tilde{h}, \tilde{j}) = \arg\max_{(h,j)\in\mathcal{L}_t} \bar{U}_{h,j}(t)$
  Compute the LCB $L_{\tilde{h},\tilde{j}}(t)$
  if $U_{\tilde{h},\tilde{j}}(t) - L_{\tilde{h},\tilde{j}}(t) \leq \beta\delta(h)^\gamma$ **then**
    expand the node, remove $(\tilde{h}, \tilde{j})$ from $\mathcal{L}_t$, add to $\mathcal{L}_t$ the K sub-cells of $\mathcal{P}_{\tilde{h},\tilde{j}}$ and sample each new node once, $n = n + K$, $t = t + 1$
  else
    Sample the state $x_t = x_{\tilde{h},\tilde{j}}$ and collect the observation $Y_{x_{h_t,j_t}}$, $n = n + K$, $t = t + 1$
  end
end
Return the node according to the returning rule.

The analysis relies on the possibility to construct, for any $\eta > 0$, upper- and lower-confidences bounds $U^\eta_{h,j}(t)$ and $L^\eta_{h,j}(t)$ such that the event

$$\mathcal{A}_\eta = \bigcap_{T\geq t\geq 1} \bigcap_{\mathcal{P}_{h,j}\in\mathcal{T}_t} \left\{ U^\eta_{h,j}(t) \geq g(x_{h,j}),\ L^\eta_{h,j}(t) \leq g(x_{h,j}) \right\}$$

has probability at least $\mathbb{P}(\mathcal{A}_\eta) \geq 1 - \eta$. We defer to Section 4 their specific expression for the case of the quantile.

Contrary to the framework of Munos et al. [2014], in our setting the magnitude of the confidence bound ($i.e$ E) associated to each node is not explicit. We thus need to introduce the following definition to quantify how many times a node needs to be sampled before satisfying the expansion condition (Eq. 2).

Definition 1 *Let*

$$n_{\eta,h}(\kappa, \alpha) = \log(T^2/\eta)\left(\frac{\kappa}{\beta\delta(h)^\gamma}\right)^\alpha \quad \textit{and } N_{h,j}(t) = \sum_{s=1}^{t} \mathbb{1}_{X(s)\in\mathcal{P}_{h,j}} \ .$$

The vector of safe constants $v = (\kappa, \alpha)$ *is composed of the constants* $\kappa' > 0$ *and* $\alpha' > 0$ *such that the event*

$$\mathcal{B}_\eta = \bigcap_{T\geq t\geq 1} \bigcap_{N_{h,j}\geq n_{\eta,h}(\kappa',\alpha')} \bigcap_{\mathcal{P}_{h,j}\in\mathcal{T}_t} \left\{ U^\eta_{h,j}(t) - L^\eta_{h,j}(t) \leq \beta\delta(h)^\gamma \right\}$$

has probability at least $1 - \eta$.

Note that in the case of the conditional expectation, Munos et al. [2014] take $\alpha = 2$, $\kappa = \sqrt{1/2}$ and $n_{\eta,h} = \dfrac{\log(T^2/\eta)}{2(\beta\delta(h)^\gamma)^2}$.

We first prove (Proposition 1) that any point at the center of an expanded cell of depth h belongs to

$$J_h = \{ \text{ } x_{h,j} \text{ such that } g(x_{h,j}) + 2\beta\delta(h)^\gamma \geq g^* \}. \tag{3}$$

Next, we show that using a budget T, the tree $\mathcal{T}_T$ reaches at least a depth $H^*_\eta(T)$ given below (Proposition 2). This implies that the point returned by the algorithm belongs to $J_{H^*_\eta(T)}$ (Proposition 3). Finally, using an assumption on the size of J_h that can be formalized by the so-call *near-optimality dimension* [Bubeck et al., 2011, Munos et al., 2014], we provide an upper bound on the regret (Theorem 1).

Proposition 1 *Conditionally on* $\mathcal{A}_\eta$*, StoROO only expands cells* $\mathcal{P}_{h,j}$ *such that* $x_{h,j} \in J_h$.

Given the value $n_{\eta,h}$ and the total budget T, the deeper the algorithm builds the tree, the better are the guarantees on the final point returned. So the goal of the following proposition is to provide a lower bound on the depth of $\mathcal{T}_T$.

Proposition 2 *Define H_η the largest $h \in \mathbb{N}$ such that*

$$S_h = K \sum_{h' \leq h} n_{\eta,h'+1} |J_{h'}| \leq T,$$

with $|J_{h'}|$ the cardinal of $J_{h'}$. The deepest node H_η^* expanded by StoROO is such that $H_\eta^* \geq H_\eta$.

Intuitively, S_h is the budget needed to expand all the nodes in J_h for all $h' \leq h$. It may be that some of this nodes will not be visited, but in the worst case they are and they need to be considered in order to obtain a valid bound. Putting Propositions 1 and 2 together, yields a first upper bound on the simple regret:

Proposition 3 *Running StoROO with budget T, with probability $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta)$ the regret is bounded as*

$$r_T \leq 2\beta\delta(H_\eta^*(T))^\gamma.$$

A more explicit bound for the regret can be obtained by quantifying the volume of

$$\mathcal{X}_\epsilon = \{x \in \mathcal{X},\ f(x) \leq f^* - \epsilon\}$$

for small values of ϵ. Introducing the Holderian semi-metric

$$l_{\beta,\gamma}(x, x') = \beta \left\| x - x' \right\|^\gamma,$$

that is associated with its regularity constants β and γ, the *near-optimality* dimension of the function is defined as follows, see Munos et al. [2014], Bubeck et al. [2011].

Definition 2 *The ν-near optimality dimension is the smallest $d \geq 0$ such that for all $\epsilon \geq 0$, there exists $C \geq 0$ such that the maximal number of disjoint $l_{\beta,\gamma}$-balls of radius $\nu\epsilon$ with center in $\mathcal{X}_\epsilon$ is less than $C\epsilon^{-d}$.*

To evaluate H_η^* we need to bound $|J_h|$ for all $h \geq 0$. The following proposition makes the link between the near optimality dimension and $|J_h|$.

Proposition 4 *Let d be the $\frac{\nu^\gamma}{2}$-near-optimality dimension, and C the corresponding constant. Then*

$$|J_h| \leq \frac{C}{\left(2\beta\delta(h)^\gamma\right)^d}.$$

Finally, combining Propositions 3 and 4 with an hypothesis on the decreasing sequence $\delta(h)$, it is possible to provide the speed of convergence of r_T.

Theorem 1 *Assume that $\delta(h) = c\rho^h$ for some $c \geq 0$ and $\rho < 1$, and assume that $\upsilon = (\kappa, \alpha)$. Thus with probability $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta)$, the regret of StoOO is bounded as*

$$r_T \leq c_1 \Big[\frac{\log(T^2/\eta)}{T}\Big]^{\frac{1}{\alpha+d}}, \quad \textit{with } c_1 = 2\beta\left[\frac{KC\kappa^\alpha[2\beta]^{-d}}{(1-\rho^{d\gamma+\gamma\alpha})}\right]^{\frac{1}{d+\alpha}},$$

where d is the near optimality dimension and C the corresponding near optimality constant.

Remark: In the particular case where each cell is a hypercube and the sub-regions are created by the division of the parent-cell into $K = 2^D$ sub-regions of equal size, then $K = 2^D$, c is equal to $\sqrt{D}$ and ρ is equal to $\frac{1}{2}$.

4 Optimizing Quantiles

In this section, we focus on the optimization of *quantiles*, which are well-established tools in (risk-averse) decision theory [see Rostek, 2010, for instance]. In particular, they benefit from interesting robustness properties, with respect to outliers or heavy tails. Let

$$g(x) = q_x(\tau) = \inf\big\{q \in \mathbb{R} : F_x(q) \geq \tau\big\},$$

now denote the τ-quantile of Y_x, where F_x is the cumulative distribution function (CDF) of $\mathbb{P}_x$.

In this section we detail how to construct the UCB and LCB for quantiles. First, we provide bounds based on Hoeffding's inequality and we use them to adapt the regret bounds of Theorem 3. Then we provide two more refined

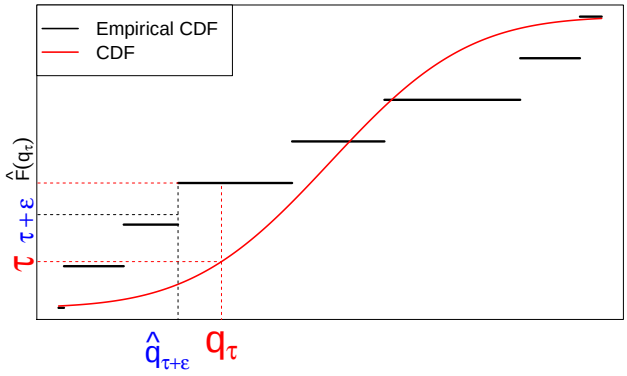

Figure 1: Illustration of the equivalence (4).

bounds that take into account the order τ of the quantile based respectively on the Bernstein's inequality and on the Kullback-Leibler divergence.

Let us first introduce some notation. For all $1 \leq t \leq T$, $1 \leq h \leq t$, $1 \leq j \leq K^h$ and $q \in \mathbb{R}$ we denote

$$\hat{F}^t_{h,j}(q) = \frac{\sum_{s=1}^t \mathbb{1}_{Y(t)\leq q} \mathbb{1}_{X(t)\in\mathcal{P}_{h,j}}}{N_{h,j}(t)}$$

the empirical CDF of the reward inside the cell $\mathcal{P}_{h,j}$, where $N_{h,j}(t)$ is the (random) number of times the cell was sampled up to time t (see Definition 1). The *generalized inverse* $\hat{F}^{t\ -}_{h,j}$ of the piecewise constant function $\hat{F}^t_{h,j}$ is defined as

$$\hat{q}_{h,j}(\tau) = \inf\left\{q \in \mathbb{R} :\ \hat{F}^t_{h,j}(q) \geq \tau\right\},$$

that is the $\lceil N_{h,j}(t) \times \tau \rceil$ order statistic of the sample that has been collected from the node $x_{h,j}$ until time t.

To define confidence bounds on the conditional quantile we proceed in two steps. First we propose confidence bounds on $\hat{F}_{h,j}(q_\tau)$. To do so, we simply use deviation bounds for Bernoulli distributions, since for all $x \in \mathcal{X}$, for all $1 \leq n \leq T$, the random variables $\left(\mathbb{1}_{Y_x(\xi_s)\leq q_x(\tau)}\right)_{s=1,\cdots,n}$ are independent and identically distributed with a Bernoulli law of parameter τ, if ξ_s denotes the time when the node x has been sampled for the s-th time. Then we use the properties

$$\forall\ \epsilon > 0\ \text{ such that }\ \tau + \epsilon < 1, \quad \hat{F}^t_{h,j}(q_{h,j}(\tau)) \geq \tau + \epsilon \quad \Leftrightarrow \quad q_{h,j}(\tau) \geq \hat{F}^{t\ -}_{h,j}(\tau + \epsilon), \tag{4}$$

$$\forall\ \epsilon > 0\ \text{ such that }\ \tau + \epsilon > 0, \quad \hat{F}^t_{h,j}(q_{h,j}(\tau)) < \tau - \epsilon \quad \Leftrightarrow \quad q_{h,j}(\tau) \leq \hat{F}^{t\ -}_{h,j}(\tau - \epsilon), \tag{5}$$

to create confidence bounds on $q_{h,j}(\tau)$ using bounds on $\hat{F}^t_{h,j}(q_\tau)$. The first equivalence in illustrated on Figure 1.

4.1 Hoeffding's bound and regret analysis

Let $\epsilon^{\eta,T}_{N_{h,j}(t)} = \sqrt{\dfrac{\log(2T^2/\eta)}{2N_{h,j}(t)}}$, and let

$$U^\eta_{h,j}(t) = \begin{cases} \inf\left\{q,\ \hat{F}^t_{h,j}(q) \geq \tau + \epsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau + \epsilon^{\eta,T}_{N_{h,j}(t)} < 1 \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

$$L^\eta_{h,j}(t) = \begin{cases} \max\left\{q,\ \hat{F}^t_{h,j}(q) \geq \tau - \epsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau - \epsilon^{\eta,T}_{N_{h,j}(t)} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The next proposition motivates the choice of the above quantities as a UCB and a LCB for the quantile of order τ at the points $(x_{h,j})_{(h,j)\in\mathcal{T}_t}$.

Proposition 5 *For any $\eta > 0$, for all $h \geq 0$, for all $0 \leq j \leq K^h$ and for all $1 \leq t \leq T$, if $L^{\eta}_{h,j}(t)$ and $U^{\eta}_{h,j}(t)$ are defined according to (7) and (6), respectively, then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.*

Now, analyzing the regret requires a high probability bound on the number of time a node is sampled before being expanded:

Proposition 6 *Assuming that for all $x \in \mathcal{X}$, Y_x has a density f_x supported in $[a,b]$, $0 \leq a \leq b \leq 1$, such that $\bar{f}(x) = \min_{y\in[a,b]} f_x(y) > 0$. If $L^{\eta}_{h,j}(t)$ and $U^{\eta}_{h,j}(t)$ are defined according to (7) and (6), respectively then a vector of safe constants is given as*

$$v = \left(\frac{2\sqrt{2}}{\inf_{x\in\mathcal{X}} \bar{f}(x)}, 2\right),$$

and for any $\eta > 0$

$$\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta) \geq 1-\eta.$$

According to the previous proposition, if we have sampled a node at depth h more than

$$n_{\eta,h}(\kappa,\alpha) = \log(T^2/\eta)\left(\frac{2\sqrt{2}}{\min_{x\in\mathcal{X}} \bar{f}(x)\beta\delta(h)^\gamma}\right)^2 \tag{8}$$

times, then with probability $1-\eta$ Condition (2) is satisfied and thus the node is expanded.

Equation (8) reflects that the smaller the minimum (taken over the whole support) of the density, the larger the upper bound on the number of samples needed before being expanded. Actually the bound is crude. It is rather clear, in fact, that the *local* minimum of f_x around $q_x(\tau)$ is the crucial quantity. Here we chose to write the results in terms of the global minimum to simplify the proof of Proposition (6). A more precise way to understand the behaviour of StoROO is that the number of time a node needs to be sampled before expansion depends on the pdf value in a neighborhood (of decreasing size with N) of the targeted quantile.

To obtain an upper bound on the simple regret, we now just need to combine Theorem 1 with Proposition 6 that provides the following theorem.

Theorem 2 *Assume that $\delta(h) = c\rho^h$ for some $c \geq 0$ and $\rho < 1$, then with probability $1-\eta$, the regret of StoROO for minimizing the quantile is bounded as*

$$r_T \leq c_2\left[\frac{\log(T^2/\eta)}{T}\right]^{\frac{1}{2+d}} \text{ with } c_2 = 2\beta\left[\frac{8KC[2\beta]^{-d}}{\min_{x\in\mathcal{X}} \bar{f}(x)^2(1-\rho^{d\gamma+\gamma\alpha})}\right]^{\frac{1}{d+2}},$$

with d the near-optimality dimension and C the near-optimality corresponding constant.

Note that the speed of convergence is the same as the one obtained in the conditional expectation optimization setting; only the constant varies.

4.2 Tight bounds

Using Hoeffding's inequality is convenient because it leads to explicit lower and upper confidence bounds, which simplifies the deriviation of bounds on the regret. However, it implicitly upper-bounds the variance of all $[0,1]$-valued random variables by $1/4$, which is overly pessimistic when the inequality is applied to variables whose expectations are far from $1/2$. This is in particular the case for quantile estimation, when the quantile is of order close to 0 or 1. To take into account the order of the quantile, following David and Shimkin [2016], a first possibility is to derive confidence intervals from Bernstein's inequality as presented in the following theorem.

Proposition 7 *For any $\eta > 0$, for all $1 \leq t \leq T$, $1 \leq h \leq t$ and $1 \leq j \leq K^h$, define*

$$U^{\eta}_{h,j}(t) = \begin{cases} \min\left\{q,\ \hat{F}^t_{h,j}(q) \geq \tau + \epsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau + \epsilon^{\eta,T}_{N_{h,j}(t)} < 1 \\ 1 & \text{otherwise,} \end{cases}$$

and

$$L^{\eta}_{h,j}(t) = \begin{cases} \max\left\{q,\ \hat{F}^t_{h,j}(q) \geq \tau - \epsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau - \epsilon^{\eta,T}_{N_{h,j}(t)} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

with

$$\epsilon^{\eta,T}_{N_{h,j}(t)} = \frac{\log(2T^2/\eta)}{3N_{h,j}(t)}\left(1 + \sqrt{1 + \frac{18N_{h,j}(t)\tau(1-\tau)}{\log(2T^2/\eta)}}\right).$$

Then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

The proof is deferred to Supplementary Material. Although Bernstein's inequality takes into account the order of the quantile, it is possible to do something better. In order to create tighter confidence bound, we thus go back to Chernoff's inequality and derive less explicit, but more accurate upper- and lower- confidence bounds on the τ-quantiles. We follow here Garivier and Cappé [2011], but a close inspection at the proofs shows however a difference in the order of the marginals of the KL functions. Recall that the binary relative entropy is defined for $(p,q) \in [0,1]^2$ as:

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q},$$

with by convention, $0\log 0 = 0$, $\log 0/0 = 0$ and $x\log x/0 = +\infty$ for $x > 0$.

Proposition 8 *For any $\eta > 0$, for all $1 \le t \le T$, $1 \le h \le t$ and $1 \le j \le K^h$, define*

$$U^{\eta}_{h,j}(t) = \min\Big\{q,\ \hat{F}^n_{h,j}(q) \ge \tau \text{ and } N_{h,j}(t)\,\mathrm{kl}(\hat{F}^t_{h,j}(q),\tau) \ge \log\frac{2T^2}{\eta}\Big\}$$

if

$$N_{h,j}(t)\,\mathrm{kl}(1,\tau) > \log\frac{2T^2}{\eta} \quad \text{and} \quad 1 \quad \text{otherwise.}$$

Define

$$L^{\eta}_{h,j}(t) = \max\Big\{q,\ \hat{F}^t_{h,j}(q) \le \tau \text{ and } N_{h,j}(t)\,\mathrm{kl}(\hat{F}^n_{h,j}(q),\tau) \ge \log\frac{2T^2}{\eta}\Big\}$$

if

$$N_{h,j}(t)\,\mathrm{kl}(0,\tau) > \log\frac{2T^2}{\eta} \quad \text{and} \quad 0 \quad \text{otherwise.}$$

Then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

Contrary to Bernstein's inequality, Chernoff's bound $\mathbb{P}\big(\hat{F}^n(q(\tau)) \ge x\big) \le \exp(-n\,\mathrm{kl}(x,\tau))$ is always tighter than Hoeffding's inequality $\mathbb{P}(\hat{F}^n\big(q(\tau)\big) \ge x\big) \le \exp\big(-2n(\tau-x)^2\big)$, which follows from Pinsker's inequality [see e.g. Garivier et al., 2018]:

$$\forall\, 0 \le p < q \le 1,\ \mathrm{kl}(p,q) \ge \frac{1}{2\max_{x\in[p,q]} x(1-x)}(p-q)^2 \ge 2(p-q)^2\ .$$

For example, given $\tau > 0.5$ and an i.i.d. sample of size n, one can see that

$$U^{\mathrm{kl}}_n \le \hat{q}_n\left(\tau + \sqrt{\frac{2\tau(1-\tau)\log(2/\eta)}{n}}\right) < \hat{q}_n\left(\tau + \sqrt{\frac{\log(2/\eta)}{2n}}\right) = U^{\mathrm{H}}_n,$$

with U^{kl} (resp. U^{H}) the UCB associated to Chernoff's inequality (resp. Hoeffding's inequality). Berstein's inequality is tighter than Hoeffding's when τ is different from $1/2$ and n sufficiently large, but always looser than Chernoff. It follows in particular that the regret of StoROO using confidence bounds derived from Chernoff's inequality has, at least, the guarantees presented in Theorem 5.

The online setting we consider in this article induces that, after t steps, the set of nodes and the number of observations in each node are random. To cope with this, we thus need deviation bounds for random size samples. The most simple way to obtain such inequalities is to use a union bound on the possible number of observations in each node, as presented above. Tighter results can be obtained from a more thorough analysis (sometimes called *peeling trick*): this is what is presented below.

Proposition 9 *For any $\eta > 0$ let*

$$\delta_\eta(T) = \inf\big\{\delta,\ Te\lceil\delta\log(T)\rceil\exp(-\delta) \le \eta/2\big\},$$

and define

$$U^{\eta}_{h,j}(t) = \min\Big\{q,\ \hat{F}^n_{h,j}(q) \ge \tau \text{ and } N_{h,j}(t)\,\mathrm{kl}(\hat{F}^t_{h,j}(q),\tau) \ge \delta_\eta(T)\Big\}$$

if

$$N_{h,j}(t)\,\mathrm{kl}(1,\tau) > \delta_\eta(T) \quad \text{and} \quad 1 \quad \text{otherwise.}$$

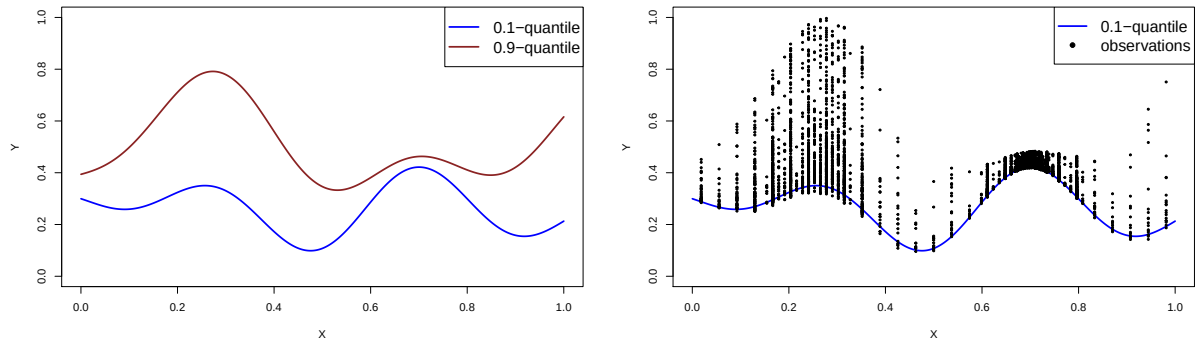

Figure 2: To the left: conditional quantiles of Φ, to the right: one run of $\mathrm{StoROO_{kl}}$ for the optimization of the 0.1-quantile with $T = 5000$, $\beta = 12$ and $\gamma = 1.4$.

Define

$$L^{\eta}_{h,j}(t) = \max\left\{q,\ \hat{F}^{t}_{h,j}(q) \leq \tau \text{ and } N_{h,j}(t)\,\mathrm{kl}(\hat{F}^{n}_{h,j}(q), \tau) \geq \delta_{\eta}(T)\right\}$$

if

$$N_{h,j}(t)\,\mathrm{kl}(0,\tau) > \delta_{\eta}(T) \quad \textit{and} \quad 0 \quad \textit{otherwise.}$$

Then the event $\mathcal{A}_{\eta}$ has probability at least $1-\eta$.

5 Experiments

We empirically highlight the capacity of StoROO to optimize the conditional quantile of a black-box function. Four versions of StoROO are compared, $\mathrm{StoROO_{H}}$ ($i.e$ StoROO using confidence bounds derived from Hoeffding's inequality), $\mathrm{StoROO_{B}}$ ($i.e$ StoROO using confidence bounds derived from Bernstein's inequality), $\mathrm{StoROO_{kl}}$ ($i.e$ StoROO using confidence bounds derived from Chernoff's inequality) and $\mathrm{StoROO_{kl\text{-}p}}$ ($i.e$ StoROO using confidence bounds derived from Chernoff's inequality and the *peeling trick*).

As a test-case, we use the function

$$\Phi(x) = \frac{0.3(\sin(3x - 0.3)\sin(13x - 1.3)) + 1.3 + 0.1\zeta(\cos(8x - 2.4) + 1.2)}{1.63},$$

with ζ following a log-normal distribution of parameter 0 and 1 truncated at its 0.95-quantile with the truncated mass following a uniform distribution between 3.85 and 5.18. Figure 5 (left) shows the shape of the 0.1 and 0.9 quantiles of g, while Figure 5 (right) shows samples of g.

The performance of each version of StoROO is evaluated for different values of τ and quantified according to the simple regret. In our experiments we fix the values $\beta = 12$ and $\gamma = 1.4$ such that the condition (1) is satisfied. Note that these values do not correspond to the actual regularity conditions at optimum. In addition we fix $K = 3$ and we choose to expand the nodes into three sub-region of equal sizes.

Figure 5 reports the average of the simple regret over 1000 runs for $\tau = 0.1$ and $\tau = 0.9$. For both values of τ all the variants of StoROO have a regret that decreases with the budget. However from our experiments a ranking can be created.

The less efficient method is $\mathrm{StoROO_{H}}$. For $\tau = 0.9$ its simple regret decreases slower than the three others methods and for $\tau = 0.1$ $\mathrm{StoROO_{H}}$ does not reach the performance of the others variants. Sometimes to reach a fixed accuracy, $\mathrm{StoROO_{H}}$ needs a much larger budget than others variants. For example taking $\tau = 0.9$, $\mathrm{StoROO_{H}}$ needs a budget of 15000 to reach a simple regret of order 10^{-4}, while $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$ need a budget equals to 5000.

Then there is $\mathrm{StoROO_{B}}$. Using the maximal budget, on both experiments this variant reaches the same accuracy as $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$ but its simple regret decreases slower. For some levels of performance $\mathrm{StoROO_{B}}$ needs a much larger budget than $\mathrm{StoROO_{kl}}$. For example, taking $\tau = 0.1$, to reach the value $r_T = 1 \times 10^{-4}$ $\mathrm{StoROO_{B}}$ needs the budget $T = 15000$ while $T = 10000$ is enough for $\mathrm{StoROO_{kl}}$.

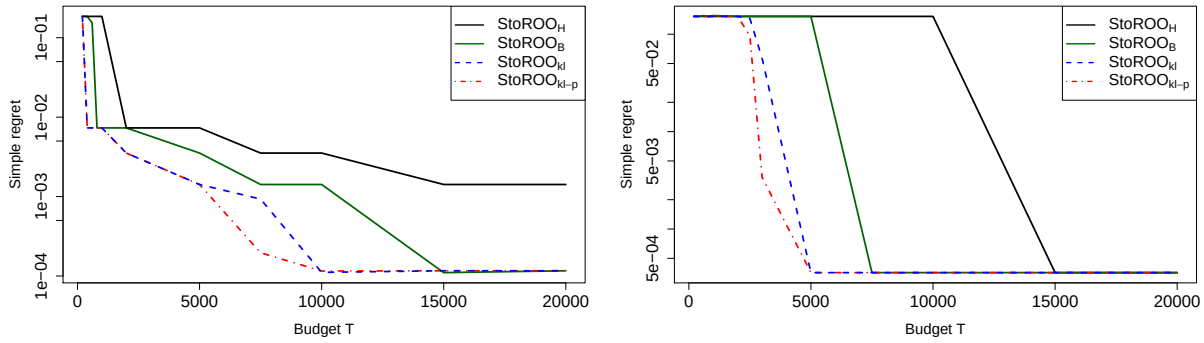

Figure 3: Evolution of the expectation of the simple regret for the optimization of the conditional quantile of Φ: to the left $\tau = 0.1$, to the right $\tau = 0.9$.

Finally, the most efficient methods are $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$. Both methods are always better or equal to $\mathrm{StoROO_{H}}$ and $\mathrm{StoROO_{B}}$. The variants $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$ are often equivalent but sometimes the regret of $\mathrm{StoROO_{kl\text{-}p}}$ decreases slightly faster than the version without the peeling trick. This behaviour provides a small gain for $\mathrm{StoROO_{kl\text{-}p}}$.

6 Conclusion

In this work, we extended StoSOO to a generic algorithm applicable to any functional of the reward distribution. We proposed a tailored application to the problem of quantile optimization, with four variants: one based on the classical Hoeffding's inequality, one based on Bernstein's inequality, and two others based on Chernoff's inequality. We showed that using Chernoff's inequality to build confidence intervals resulted in a dramatic improvement, both in theory and practice.

For simplicity, we assumed in this paper that the local regularity (or at least, an upper bound) of the target function at the optimum was known to the user. However, we believe that it is possible to combine our results to the procedure defined in Grill et al. [2015], Xuedong et al. [2019] so that creating an algorithm able to optimize g without the knowledge of the smoothness near an optimal point: this is left for future work. A second possible extension is to leverage the results proposed here to design an algorithm for the cumulative regret, in the spirit of HOO Bubeck et al. [2011] for example.

References

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. Online stochastic optimization under correlated bandit feedback. In *ICML*, pages 1557–1565, 2014.

Peter L Bartlett, Victor Gabillon, and Michal Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. *arXiv preprint arXiv:1810.00997*, 2018.

Fabio Bellini and Elena Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6): 487–506, 2017.

Clément Bouttier. Optimisation globale sous incertitudes: algorithmes stochastiques et bandits continus avec application à la planification de trajectoires dâavions. 2017.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

Yahel David and Nahum Shimkin. Pure exploration for max-quantile bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 556–571. Springer, 2016.

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018.

Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems*, pages 667–675, 2015.

Adam J Hepworth. *A multi-armed bandit approach to superquantile selection*. PhD thesis, Monterey, California: Naval Postgraduate School, 2017.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

Ravi Kumar Kolla, Krishna Jagannathan, et al. Risk-aware multi-armed bandits using conditional value-at-risk. *arXiv preprint arXiv:1901.00997*, 2019.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492, 2018.

Alexander J McNeil and Rüdiger Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3-4):271–300, 2000.

Rémi Munos et al. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Marzena Rostek. Quantile maximization in decision theory. *The Review of Economic Studies*, 77(1):339–371, 2010.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

Xuedong Shang, Emilie Kaufmann, and Michal Valko. General parallel optimization without a metric. In *30th International Conference on Algorithmic Learning Theory*, 2019.

Balazs Szorenyi, Róbert Busa-Fekete, Paul Weng, and Eyke Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *32nd International Conference on Machine Learning*, pages 1660–1668, 2015.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27, 2013.

Shang Xuedong, Emilie Kaufmann, and Michal Valko. General parallel optimization a without metric. In *Algorithmic Learning Theory*, pages 762–787, 2019.

A Proofs related to the generic analysis of StoROO

Proof of Proposition 1

Let us define $\mathcal{P}_{h^*,j^*}$ the partition containing x^*. Assume that the partition $\mathcal{P}_{h,j}$ has been selected, thus

$$\bar{U}_\eta^{h,j}(t) \geq \bar{U}_\eta^{h^*,j^*}(t).$$

By definition $\bar{U}_\eta^{h^*,j^*}(t) \geq f^*$, thus $\bar{U}_\eta^{h,j}(t) \geq f^*$. Conditionally on $\mathcal{A}_\eta$, $L_\eta^{h,j}(t)) \leq f(x_{h,j}(t))$ that implies

$$f^* - f(x_{h,j}) \leq \bar{U}_\eta^{h,j}(t) - L_\eta^{h,j}(t) \leq U_\eta^{h,j}(t) + \beta\delta(h)^\gamma - L_\eta^{h,j}(t) \leq 2\beta\delta(h)^\gamma.$$

Note that the last inequality is obtained because the partition is expanded, which implies that

$$U(x_{h,j})(t) - L(x_{h,j})(t) \leq \beta\delta(h)^\gamma.$$

Finally:

$$f^* \leq f(x_{h,j}) + 2\beta\delta(h)^\gamma,$$

thus $x_{h,j}$ belongs to J_h.

Proof of Proposition 2

$$\begin{aligned}
T &= \sum_{h,j \in \mathcal{T}_T} N_{h,j}(t) \leq \sum_{h,j \in \mathcal{T}_T} n_{\eta,h} \quad \text{because } N_{h,j}(t) \leq n_{\eta,h} \\
&\leq \sum_{h'=0}^{\text{depth}(\mathcal{T}_T)-1} K|\mathcal{T}_T \cap J_h| n_{\eta,h'+1} \quad \text{because StoROO has not expanded all the nodes it has sampled} \\
&\leq \sum_{h'=0}^{\text{depth}(\mathcal{T}_T)-1} K|J_h| n_{\eta,h'+1} = S_{\text{depth}(\mathcal{T}_T)-1}.
\end{aligned}$$

Thus $S_{H_\eta} \leq S_{\text{depth}(\mathcal{T}_T)-1} \leq S_{\text{depth}(\mathcal{T}_T)}$ so $H_\eta \leq \text{depth}(\mathcal{T}_T)$. There is at least an expanded node of depth $H^*_\eta \geq H_\eta$ after a budget T was used.

Proof of Proposition 4 According to the assumption 2, each cell $\mathcal{P}_{h,j}$ contains ball of radius $\nu\delta(h)$ centered in $x_{h,j}$ that is a $l_{\beta,\gamma}$ ball of radius $\beta(\nu\delta(h))^\gamma$ centered in $x_{h,j}$. If the d is the $\nu^\gamma/2$ near optimality dimension then there is at most $C[2\beta\delta(h)^\gamma]^{-d}$ disjoint $l_{\beta,\gamma}$ balls of radius $\beta(\nu\delta(h))^\gamma$ inside $\mathcal{X}_{2\beta\delta(h)^\gamma}$. Thus if $|J_h| = |x_{h,j} \in \mathcal{X}_{2\beta\delta(h)^\gamma}| > C[2\beta(\delta(h))^\gamma]^{-d}$ this implies there is more than $C[2\beta\delta(h)^\gamma]^{-d}$ disjoint $l_{\beta,\gamma}$ balls of radius $\beta(\nu\delta(h))^\gamma$ with center in $\mathcal{X}_{2\beta\delta(h)^\gamma}$, that is a contradiction.

Proof of Therorem 1

$$\begin{aligned}
T &\leq \sum_{h=0}^{H^*} K|J_h| n_{\eta,h+1} \qquad \text{by definition of } H^* \\
&\leq \sum_{h=0}^{H^*} KC[2\beta\delta(h)^\gamma]^{-d} n_{\eta,h+1} \qquad \text{using Proposition 4} \\
&= \sum_{h=0}^{H^*} KC[2\beta(c\rho^h)^\gamma]^{-d} n_{\eta,h+1} \qquad \text{using the hypothesis on the exponential decay of the diameter of the cells} \\
&\leq \sum_{h=0}^{H^*} KC[2\beta(c\rho^h)^\gamma]^{-d} \times \kappa^\alpha \frac{\log(T^2/\eta)}{(\beta(c\rho^h)^\gamma)^\alpha} \qquad \text{applying Definition 1} \\
&= \log(T^2/\eta) \frac{KC\kappa^\alpha[2\beta c^\gamma]^{-d}}{\beta c^{\gamma\alpha}} \sum_{h=0}^{H^*} \rho^{h(-d\gamma-\gamma\alpha)} \\
&= \log(T^2/\eta) \frac{KC\kappa^\alpha[2\beta c^\gamma]^{-d}}{\beta c^{\gamma\alpha}} \times \frac{\rho^{(H^*+1)(-d\gamma-\gamma\alpha)} - 1}{\rho^{-d\gamma-\gamma\alpha} - 1} \qquad \text{rewritting the sum} \\
&\leq \frac{\log(T^2/\eta)}{(1-\rho^{d\gamma+\gamma\alpha})} \frac{KC\kappa^\alpha[2\beta c^\gamma]^{-d}}{\beta c^{\gamma\alpha}} \times \rho^{H^*(-d\gamma-\gamma\alpha)} \\
&= \frac{\log(T^2/\eta)}{(1-\rho^{d\gamma+\gamma\alpha})} \frac{KC\kappa^\alpha[2\beta]^{-d}}{\beta} \times \delta(H^*)^{-d\gamma-\gamma\alpha}.
\end{aligned}$$

Finally

$$\left[\frac{KC\kappa^\alpha[2\beta]^{-d}}{\beta(1-\rho^{d\gamma+\gamma\alpha})}\right]^{\frac{1}{d\gamma+\gamma\alpha}} \left[\frac{\log(T^2/\eta)}{T}\right]^{\frac{1}{d\gamma+\gamma\alpha}} \geq \delta(H^*).$$

Using Proposition 3 we obtain

$$r_T \leq c_1 \left[\frac{\log(T^2/\eta)}{T}\right]^{\frac{1}{\alpha+d}}.$$

B Proofs related to the section Optimizing quantiles

Proof of Proposition 5

Let us consider the event

$$\begin{aligned}\xi_\eta &= \{\forall\, h \geq 0, \forall\, 0 \leq j \leq K^h, \forall\, 1 \leq t \leq T, \\ &\quad \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^\eta_{N_{h,j}(t)} \text{ or } \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) < \tau - \epsilon^\eta_{N_{h,j}(t)}\}.\end{aligned}$$

$$\begin{aligned}\mathbb{P}\big(\xi_\eta\big) &= \mathbb{P}\bigg(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T, \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^\eta_{N_{h,j}(t)} \text{ or },\\ &\qquad\qquad \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) < \tau - \epsilon^\eta_{N_{h,j}(t)}\bigg)\\ &\leq \mathbb{P}\bigg(\forall h \leq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T, \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^\eta_{N_{h,j}(t)})\bigg)\\ &\quad + \mathbb{P}\bigg(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T, \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) < \tau - \epsilon^\eta_{N_{h,j}(t)}\bigg)\end{aligned}$$

Define $m \leq T$ the number of nodes expanded throughout the algorithm, define for $1 \leq w \leq m$, ζ^s_w as the time when the cell w has been selected for the s-th time and define $Y_w(\zeta^s_w)$ the reward obtained at that time at the point x_w. Then one can write

$$\mathbb{P}\bigg(\hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^{\eta,T}_{N_{h,j}(t)}\bigg) = \mathbb{P}\bigg(\frac{1}{N_{h,j}(t)}\sum_{s=1}^{N_{h,j}(t)} \mathbb{1}_{Y_{h,j}(\zeta^s_{h,j}) \leq q_{h,j}(\tau)} \geq \tau + \epsilon^\eta_{N_{h,j}(t)}\bigg).$$

Using this notation, we have:

$$\begin{aligned}&\mathbb{P}\bigg(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T,\ \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^\eta_{N_{h,j}(t)}\bigg)\\ \leq\ &\mathbb{P}\Big(\exists\, 1 \leq w \leq T,\ \exists\, 1 \leq u \leq T,\ \frac{1}{u}\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leq q_w(\tau)} \geq \tau + \epsilon^\eta_u\Big)\\ \leq\ &\sum_{w=1}^{T}\sum_{u=1}^{T} \mathbb{P}\Big(\frac{1}{u}\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leq q_w(\tau)} \geq \tau + \epsilon^\eta_u\Big)\end{aligned}$$

By Hoeffding's inequality, if

$$\epsilon^\eta_u = \sqrt{\frac{\log(2T^2/\eta)}{2u}},$$

we obtain

$$\mathbb{P}\bigg(\forall h \leq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T,\ \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^\eta_{N_{h,j}(t)}\bigg) \leq \frac{\eta}{2}.$$

Now using Equation (4) we can express this inequality directly in terms of quantiles:

$$\mathbb{P}\bigg(\forall h \leq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T,\ q_{h,j}(\tau) \geq U^\eta_{h,j}(t)\bigg) \leq \frac{\eta}{2}.$$

Using the same scheme of proof with Inequality (5), we obtain:

$$\mathbb{P}\bigg(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T,\ q_{h,j}(\tau) \leq L^\eta_{h,j}(t)\bigg) \leq \frac{\eta}{2},$$

and hence $\mathbb{P}\big(\mathcal{A}_\eta\big) = 1 - \mathbb{P}\big(\xi_\eta\big) \geq 1 - \eta$.

Proof of Proposition 6

Define first the event

$$\mathcal{C}_\eta = \bigcap_{T \geq t \geq 1} \bigcap_{\mathcal{P}_{h,j} \in \mathcal{T}_t} \Big\{q_{h,j}\big(\tau + 2\epsilon^{\eta,T}_{N_{h,j}(t)}\big)\big) \geq U^\eta_{h,j}(t) \geq q_{h,j}(\tau) \geq L^\eta_{h,j}(t) \geq q_{h,j}\big(\tau - 2\epsilon^{\eta,T}_{N_{h,j}(t)}\big)\Big\},$$

with

$$\epsilon^{\eta,T}_{N_{h,j}(t)} = \sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}}.$$

Using equivalences (4) and (5), one can write:

$$\begin{aligned} & q_{h,j}\big(\tau + 2\epsilon^{\eta,T}_{N_{h,j}(t)}\big) \geq U^{\eta}_{h,j}(t) \geq q_{h,j}(\tau) \geq L^{\eta}_{h,j}(t) \geq q_{h,j}\big(\tau - 2\epsilon^{\eta,T}_{N_{h,j}(t)}\big) \\ & \Leftrightarrow \hat{F}(q_{h,j}(\tau + 2\epsilon^{\eta,T}_{N_{h,j}(t)})) \geq \tau + \epsilon^{\eta,T}_{N_{h,j}(t)} > \hat{F}(q_{h,j}(\tau) \geq \tau - \epsilon^{\eta,T}_{N_{h,j}(t)} > \hat{F}(q_{h,j}(\tau + 2\epsilon^{\eta,T}_{N_{h,j}(t)})). \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{P}(\mathcal{C}_\eta) \geq \; & 1 - \mathbb{P}(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T, \sup_{y = q_\tau, q_{\tau+\epsilon^{\eta,T}_{N_{h,j}(t)}}} |F_{h,j}(y) - \hat{F}^t_{h,j}(y)| \geq \epsilon^{\eta,T}_{N_{h,j}(t)}\,) \\ \geq \; & 1 - \mathbb{P}(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall 1 \leq t \leq T, \sup_{y \in [0,1]} |F_{h,j}(y) - \hat{F}^t_{h,j}(y)| \geq \epsilon^{\eta,T}_{N_{h,j}(t)}\,). \end{aligned}$$

Using the same notation as in the proof of Proposition 5, one can write

$$\geq \; 1 - \sum_{w=1}^{T}\sum_{u=1}^{T} \mathbb{P}(\sup_{y\in[0,1]} |F_w(y) - \frac{1}{u}\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leq q_w(\tau)}| \geq \epsilon^{\eta,T}_u).$$

Now by applying the Massart's inequality to bound

$$\mathbb{P}(\sup_{y\in[0,1]} |F_w(y) - \sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leq q_w(\tau)}| \geq \epsilon^{\eta,T}_u),$$

one obtain $\mathbb{P}(\mathcal{C}_\eta) \geq \; 1 - \eta$. Thus with probability $1-\eta$, we have:

$$U^{\eta}_{h,j}(t) - L^{\eta}_{h,j}(t) \leq q_{h,j}\big(\tau + 2\epsilon^{\eta,T}_{N_{h,j}(t)}\big) - q_{h,j}\big(\tau - 2\epsilon^{\eta,T}_{N_{h,j}(t)}\big). \tag{9}$$

Assuming that $q_{h,j}$ is differentiable in τ, by the mean value theorem, we deduce

$$q_{h,j}(\tau + 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}}) - q_{h,j}(\tau - 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}}) \leq 4\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} \max_{\tau' \in (0,1)} \frac{1}{f_{x_{h,j}} \circ F^{-1}_{x_{h,j}}(\tau')}.$$

Using (9) it is possible to write that with probability $1-\eta$:

$$U^{\eta}_{h,j} - L^{\eta}_{h,j} \leq 4\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} \frac{1}{\bar{f}_{x_{h,j}}} \leq 4\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} \frac{1}{\inf_{x\in\mathcal{X}} \bar{f}(x)}.$$

We define $n_{\eta,h}$ as the smallest n such that

$$4\sqrt{\frac{\log(2T^2/\eta)}{2n}} \frac{1}{\inf_{x\in\mathcal{X}} \bar{f}(x)} \leq \beta\delta(h)^\gamma,$$

that is

$$n_{\eta,h} = \log(T^2/\eta)\left(\frac{2\sqrt{2}}{\beta\delta(h)^\gamma \inf_{x\in\mathcal{X}} \bar{f}(x)}\right)^2.$$

To conclude, since $\mathcal{C}_\eta \subset \mathcal{A}_\eta \cap \mathcal{B}_\eta$, we obtain $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta) \geq 1 - \eta$.

Proof of Proposition 7

Let $Y_1, \cdots, Y_n$ be n *i.i.d.* random variables bounded by the interval $[0,1]$. Define $\hat{F}^n(q(\tau)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \leq q(\tau)}$. For $x > \tau$ the Bernstein's inequality gives

$$\mathbb{P}(|\hat{F}^n(q(\tau)) - \tau| > \epsilon) \leq 2\exp\left(\frac{n\epsilon^2}{2\tau(1-\tau) + 2\epsilon/3}\right).$$

Let us consider the event

$$\xi_\eta \quad = \quad \{\forall\, h \geq 0, \forall\, 0 \leq j \leq K^h, \forall\, 1 \leq t \leq T,$$

$$\hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geq \tau + \epsilon^{\eta,T}_{N_{h,j}(t)} \text{ or } \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) < \tau - \epsilon^{\eta,T}_{N_{h,j}(t)}\}.$$

Using the same lines as in Proposition 5 we have

$$\begin{aligned}\mathbb{P}(\xi_\eta) \leq & \sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(|\frac{1}{u}\sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leq q_w(\tau)} - \tau| > \epsilon^{\eta,T}_u\Big) \\ & \text{then applying Bernstein's inequality we obtain} \\ \leq & \sum_{w=1}^{T}\sum_{u=1}^{T} 2\exp\Big(-\frac{u{\epsilon^{\eta,T}_{N_{h,j}(t)}}^2}{2\tau(1-\tau)+2\epsilon^{\eta,T}_{N_{h,j}(t)}/3}\Big). \end{aligned} \tag{10}$$

By now the goal is to find $\epsilon^{\eta,T}_{N_{h,j}(t)} > 0$ such that

$$\frac{u{\epsilon^{\eta,T}_{N_{h,j}(t)}}^2}{2\tau(1-\tau)+2\epsilon^{\eta,T}_{N_{h,j}(t)}/3} = \log(2T^2/\eta).$$

Finding such $\epsilon^{\eta,T}_{N_{h,j}(t)}$ can be easily done because it is a square of a second order polynomial. The result is

$$\epsilon^{\eta,T}_{N_{h,j}(t)} = \frac{\log(2T^2/\eta)}{3u}\Big(1+\sqrt{1+\frac{18u\tau(1-\tau)}{\log(2T^2/\eta)}}\Big).$$

Plugging the value of $\epsilon^{\eta,T}_{N_{h,j}(t)}$ inside (10) concludes the proof.

Proof of Proposition 8

Step 1: bounds on $\hat{F}^n(q(\tau))$ for a iid sample

Let $Y_1, \cdots, Y_n$ be n $i.i.d.$ random variables bounded by the interval $[0,1]$. Define $\hat{F}^n(q) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\leq q}$. For $x > \tau$ Chernoff's inequality gives

$$\mathbb{P}(\hat{F}^n(q(\tau)) \geq x) \leq \exp(-n\,\mathrm{kl}(x,\tau)).$$

Let $\tau^+ > \tau$ be the value such that $\mathrm{kl}(\tau^+,\tau) = \frac{\log(2/\eta)}{n}$, then for all $x \geq \tau^+$:

$$\mathbb{P}(\hat{F}^n(q(\tau)) \geq x) \leq \mathbb{P}(\hat{F}^n(q(\tau)) \geq \tau^+) \leq \exp(n\frac{\log(2/\eta)}{n}) = \frac{\eta}{2}.$$

Now let us define the candidate for the UCB of a i.i.d sample:

$$U(n) = \min\big\{q,\ \ \hat{F}^n(q) \geq \tau \text{ and } n\,\mathrm{kl}(\hat{F}^n(q),\tau) \geq \log(2/\eta)\big\},$$

and let us remark that

$$\hat{F}^n(U(n)) \leq \hat{F}^n(q(\tau)) \Leftrightarrow \tau \leq \hat{F}^n(q(\tau)) \text{ and } \mathrm{kl}(\hat{F}^n(q(\tau)),\tau) \geq \frac{\log(2/\eta)}{n}, \tag{11}$$

thus

$$\begin{aligned}\mathbb{P}(\hat{F}^n(U(n)) \leq \hat{F}^n(q(\tau))) =& \mathbb{P}(\tau \leq \hat{F}^n(q(\tau)) \text{ and } \mathrm{kl}(\hat{F}^n(q(\tau)),\tau) \geq \frac{\log(2/\eta)}{n}) \\ \leq& \mathbb{P}(\hat{F}^n(q(\tau)) \geq \tau^+) \leq \frac{\eta}{2}.\end{aligned}$$

For $x < \tau$ let us introduce

$$L(n) = \max\big\{q,\ \ \hat{F}^n(q) \leq \tau \text{ and } n\,\mathrm{kl}(\hat{F}^n(q),\tau) \geq \log(2/\eta)\big\},$$

one proves in the same way

$$\mathbb{P}(\hat{F}^n(L(n)) > \hat{F}^n(q(\tau))) \leq \frac{\eta}{2}.$$

Step 2: Double union bound

Let us consider the event

$$\begin{aligned}\xi_\eta &= \Big\{\forall\, h\geq 0, \forall\, 0\leq j\leq K^h, \forall\, 1\leq t\leq T,\\ &\quad \hat{F}^t_{h,j}\big(q_{h,j}(\tau)\big)\geq \hat{F}^t_{h,j}(U^\eta_{h,j}) \text{ or } \hat{F}^t_{h,j}\big(q_{h,j}(\tau)\big)< \hat{F}^t_{h,j}(L^\eta_{h,j})\Big\}.\end{aligned}$$

$$\begin{aligned}\mathbb{P}\big(\xi_\eta\big)\leq\ & \mathbb{P}\Big(\forall h\leq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T, \hat{F}^t_{h,j}\big(q_{h,j}(\tau)\big)\geq \hat{F}^t_{h,j}(U^\eta_{h,j})\Big)\\ &+\mathbb{P}\Big(\forall h\geq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T, \hat{F}^t_{h,j}\big(q_{h,j}(\tau)\big)< \hat{F}^t_{h,j}(L^\eta_{h,j})\Big)\end{aligned}$$

Following the notation of the proof of Proposition 5 we have

$$\begin{aligned}&\mathbb{P}\Big(\forall h\geq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T,\ \hat{F}^t_{h,j}\big(q_{h,j}(\tau)\big)\geq \hat{F}^t_{h,j}(U^\eta_{h,j})\Big)\\ \leq\ & \mathbb{P}\Big(\exists\, 1\leq w\leq T,\ \exists\, 1\leq u\leq T,\ \sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leq q_w(\tau)}\geq \sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leq U^\eta_w}\Big)\\ \leq\ & \sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(\sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leq q_w(\tau)}\geq \sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leq U^\eta_w}\Big).\end{aligned}$$

Using the equivalence (11), the probability can be reformulated as

$$= \sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(\tau\leq \hat{F}^u(q(\tau))\ \text{ and }\ \mathrm{kl}(\hat{F}^u(q(\tau)),\tau)\geq \frac{\log(2T^2/\eta)}{u}\Big).$$

Now using the Chernoff's inequality we obtain

$$\mathbb{P}\Big(\forall h\geq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T,\ \hat{F}^t_{h,j}\big(q_{h,j}(\tau)\big)\geq \hat{F}^t_{h,j}(U^\eta_{h,j})\Big)\leq \sum_{w=1}^{T}\sum_{u=1}^{T}\exp(-u\frac{\log(2T^2/\eta)}{u}) = \eta/2.$$

By equivalence (4) this implies that, $\forall h\geq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T$, with probability at least $\eta/2$, $U^\eta_{h,j}(t)\leq q_{h,j}(\tau)$. Using the same lines one can show

$$\mathbb{P}\Big(\forall h\geq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T, \hat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)< \hat{F}^t_{h,j}(L)\Big)\leq \eta/2,$$

By equivalence (5) this implies that, $\forall h\geq 0, \forall 0\leq j\leq K^h, \forall 1\leq t\leq T$, $L^\eta_{h,j}(t)> q_{h,j}(\tau)$ with probability at least $\eta/2$,

Putting this two probabilities together prove the result.

Proof of Proposition 9

Define

$$\tilde{S}^\tau_{h,j}(n) = \sum_{i=1}^{n}\mathbb{1}_{Y_{h,j}(i)\leq q_{h,j}(\tau)}.$$

Step 1: Martingale For every $\lambda\in\mathbb{R}$, let $\phi_\tau(\lambda) = \log\mathbb{E}[\exp(\lambda\mathbb{1}_{Y_{h,j}(1)\leq q_{h,j}(\tau))}]$. Let $W^\lambda_0 = 1$ and for $n\geq 1$,

$$W^\lambda_n = \exp(\lambda\tilde{S}^\tau_{h,j}(n) - n\phi_\tau(\lambda)).$$

$(W^\lambda_n)_{n\geq 0}$ is a martingale relative to $(\mathcal{F}_n)_{n\geq 0}$. In fact,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\lambda\{\tilde{S}^\tau_{h,j}(n+1)-\tilde{S}^\tau_{h,j}(n)\}\Big)|\mathcal{F}_n\Big] =& \mathbb{E}\Big[\exp(\lambda X_{n+1})|\mathcal{F}_n\Big]\\ =& \exp\Big(\log\mathbb{E}[\exp(\lambda X_1]\Big)\end{aligned}$$

$$= \exp\Big(\{(n+1)-n\}\phi_\mu(\lambda)\Big)$$

That is equivalent to

$$\mathbb{E}\Big[\exp\Big(\lambda\{\tilde{S}^\tau_{h,j}(n+1)-\tilde{S}^\tau_{h,j}(n)\}\Big)|\mathcal{F}_n\Big] = \exp\Big(\lambda S_n - n\phi_\mu(\lambda)\Big).$$

Step 2: Peeling Let us devide the interval $\{1,\cdots,T\}$ into *slices* $\{t_{k-1}+1,\cdots,t_k\}$ of geometric increasing size. We may assume that $\delta > 1$, since otherwise the bound is trivial. Take $\xi = 1/(1-\delta_\eta(T))$, let $t_0 = 0$ and for all $k \in \mathbb{N}^*$, let $t_k = \lfloor (1+\xi)^k \rfloor$.

$$\begin{aligned}
&\mathbb{P}\Big(\forall h \geq 0, \forall 0 \leq j \leq K^h, \forall\, 1 \leq t \leq T,\ U^\eta_{h,j}(t) \leq q_{h,j}(\tau)\Big) \\
\leq\ &\mathbb{P}\Big(\exists\, h \geq 0, \exists\, 0 \leq j \leq K^h,\ \exists\, 1 \leq t \leq T,\ U^\eta_{h,j}(t) \leq q_{h,j}(\tau)\Big). \qquad (12)
\end{aligned}$$

Define $m \leq T$ the number of nodes expanded throughout the algorithm, thus for $1 \leq w \leq m$, it is possible to rewrite (12) as

$$\begin{aligned}
&\mathbb{P}\Big(\exists\, 1 \leq w \leq T,\ \exists\, 1 \leq n \leq T,\ U^\eta_w(n) \leq q_w(\tau)\Big) \\
\leq\ &\sum_{w=1}^{T} \mathbb{P}\Big(\ \exists\, 1 \leq k \leq D,\ \exists\, t_{k-1} < n \leq t_k \ \text{ and }\ U^\eta_w(n) \leq q_w(\tau)\Big) \quad \text{with } D = \frac{\log(T)}{\log(1+\eta)} \\
\leq\ &\sum_{w=1}^{T}\sum_{k=1}^{D} \mathbb{P}\Big(A_k\Big),
\end{aligned}$$

with

$$A_k = \big\{\ \exists\, t_{k-1} < n \leq t_k \ \text{ and }\ U^\eta_w(n) \leq q_w(\tau)\big\}.$$

Observe that $U^\eta_w(n) \leq q_w(\tau)$ if and only if $\frac{1}{n}\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leq U^\eta_w} \leq \frac{1}{n}\tilde{S}^\tau_w(n)$ and

$$\frac{1}{n}\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leq U^\eta_w} \leq \frac{\tilde{S}^\tau_w(n)}{n} \Leftrightarrow \tau \leq \frac{\tilde{S}^\tau_w(n)}{n} \ \text{ and } \ \mathrm{kl}(\frac{\tilde{S}^\tau_w(n)}{n}, \tau) \geq \delta_\eta(T) + \frac{1}{n}.$$

Define $\delta = \delta_\eta(T) + 1/n$, let s be the smallest integer such that $\delta/(s+1) \leq \mathrm{kl}(1,\tau)$; if $n \leq s$, then $n\,\mathrm{kl}(\frac{\tilde{S}^\tau_w(n)}{n}, \tau) \leq s\,\mathrm{kl}(\frac{\tilde{S}^\tau_w(n)}{n}, \tau) \leq s\,\mathrm{kl}(1,\tau) < \delta$ thus $\mathbb{P}(U(n) < q(\tau)) = 0$. Thus for all k such that $t_k \geq s$, we obtain $\mathbb{P}(A_k = 0)$. For k such that $t_k > s$, let $\tilde{t}_{k-1} = \max\{t_{k-1}, s\}$. Let $x \in]\tau, 1[$ be such that $\mathrm{kl}(x,\tau) = \delta/n$ and let $\lambda(x) = \log(x(1-\tau)) - \log(\tau(1-x)) > 0$, so that $\mathrm{kl}(x,\tau) = \lambda(x)x - (1-\tau+\tau\exp(\lambda(x)))$. Consider z such that $z > \tau$ and $\mathrm{kl}(z,\tau) = \delta/(1+\xi)^k$.

Observe that

- if $n > \tilde{t}_{k-1}$, then
$$\mathrm{kl}(z,\tau) = \frac{\delta}{(1+\xi)^k} \geq \frac{\delta}{(1+\xi)n};$$
- if $n \leq t_k$, then as
$$\mathrm{kl}\,\big(\frac{\tilde{S}^\tau_w(n)}{n}, \tau\big) > \frac{\delta}{n} > \frac{\delta}{(1+\xi)^k} = \mathrm{kl}(z,\tau),$$

it holds that:

$$\tau \leq \frac{\tilde{S}^\tau_w(n)}{n} \ \text{ and } \ \mathrm{kl}(\frac{\tilde{S}^\tau_w(n)}{n}, \tau) \geq \frac{\delta}{n} \Rightarrow \frac{\tilde{S}^\tau_w(n)}{n} \geq z.$$

Hence on the event $\{\tilde{t}_{k-1} < n < t_k\} \cap \{\tau \leq \frac{\tilde{S}^\tau_w(n)}{n}\} \cap \{\mathrm{kl}(\frac{\tilde{S}^\tau_w(n)}{n}, \tau) \geq \frac{\delta}{n}\}$ it holds that

$$\lambda(z)\frac{\tilde{S}^\tau_w(n)}{n} \geq \lambda(z)z - \phi_\tau(\lambda(z)) = \mathrm{kl}(z,\tau) \geq \frac{\delta}{(1+\xi)n}.$$

Step 3: Putting everything together

$$\{\tilde{t}_{k-1} < n < t_k\} \cap \{\tau \leq \frac{\tilde{S}_w^\tau(n)}{n}\} \cap \{\mathrm{kl}(\frac{\tilde{S}_w^\tau(n)}{n}, \tau) \geq \frac{\delta}{n}\}$$

$$\subset \{\lambda(z)\frac{\tilde{S}_w^\tau(n)}{n} - \phi_\tau(\lambda(z)) \geq \frac{\delta}{n(1+\xi)}\}$$
$$\subset \{\lambda(z)S_w(n) - n\phi_\tau(\lambda(z)) \geq \frac{\delta_\eta(T)}{(1+\xi)}\}$$
$$\subset \{W_n^{\lambda(z)} > \exp(\frac{\delta_\eta(T)}{(1+\xi)})\}.$$

As $(W_n^\lambda)_{n\geq 0}$ is a martingale, $\mathbb{E}[W_n^{\lambda(z)}] \leq \mathbb{E}[W_0^{\lambda(z)}] = 1$. Thus the Doob's inequality for martingales provides:

$$\mathbb{P}\left(\sup_{\tilde{t}_{k-1}<n<t_k} W_n^{\lambda(z)} > \exp\left(\frac{\delta_\eta(T)}{1+\xi}\right)\right) \leq \exp\Big(-\frac{\delta_\eta(T)}{1+\xi}\Big)$$

Finally

$$\sum_{w=1}^{T}\sum_{k=1}^{D}\mathbb{P}\Big(\exists\, t_{k-1} < n \leq t_k \text{ and } U_w^\eta(n) \leq q_w(\tau)\Big) \leq TD\exp(-\frac{\delta_\eta(T)}{(1+\xi)}).$$

But as $\xi = 1/(\delta_\eta(T)-1)$, $D = \Big\lceil \dfrac{\log(T)}{\log(1+1/(\delta_\eta(T)+1))} \Big\rceil$ and as long as

$$\log(1+1/(\delta_\eta(T)-1)) \geq 1/\delta_\eta(T),$$

we obtain:

$$\mathbb{P}(\mathcal{A}^c) \leq T\Big\lceil \frac{\log(T)}{\log(1+1/(\delta_\eta(T)+1))}\Big\rceil \exp(-\delta_\eta(T)+1) \leq Te\lceil \delta_\eta(T)\log(T)\rceil \exp(-\delta_\eta(T)) \leq \eta/2.$$

Using the same lines for the LCB concludes the proof.