A ranking approach to global optimization

Cédric Malherbe MALHERBE@CMLA.ENS-CACHAN.FR
Emile Contal CONTAL@CMLA.ENS-CACHAN.FR
Nicolas Vayatis VAYATIS@CMLA.ENS-CACHAN.FR
CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235, Cachan, France

Abstract

We consider the problem of maximizing an *unknown* function f over a compact and convex set $\mathcal{X} \subset \mathbb{R}^d$ using as few observations $f(x)$ as possible. We observe that the optimization of the function f essentially relies on learning the induced bipartite ranking rule of f. Based on this idea, we relate global optimization to bipartite ranking which allows to address problems with high dimensional input space, as well as cases of functions with weak regularity properties. The paper introduces novel meta-algorithms for global optimization which rely on the choice of any bipartite ranking method. Theoretical properties are provided as well as convergence guarantees and equivalences between various optimization methods are obtained as a by-product. Eventually, numerical evidence is given to show that the main algorithm of the paper which adapts empirically to the underlying ranking structure essentially outperforms existing state-of-the-art global optimization algorithms in typical benchmarks.

1. Introduction

In many applications such as complex system design or hyperparameter calibration for learning systems, the goal is to optimize some output value of a non-explicit function with as few evaluations as possible. Indeed, in such contexts, one has access to the function values only through numerical evaluations by simulation or cross-validation with significant computational cost. Moreover, the operational constraints generally impose a sequential exploration of the solution space with small samples. The generic problem of sequentially optimizing the output of an unknown and potentially *non-convex* function is often referred to as *global optimization* (Pintér, 1991), black-box optimization (Jones et al., 1998) or derivative-free optimization (Rios & Sahinidis, 2013). There are several algorithms based on various heuristics which have been introduced in order to address complicated optimization problems with limited regularity assumptions, such as genetic algorithms, model-based algorithms, branch-and-bound methods... (see (Rios & Sahinidis, 2013) for a recent overview). This paper follows the line of the approaches recently considered in the machine learning literature (Bull, 2011; Munos, 2011; Sergeyev et al., 2013). These approaches extend the seminal work on Lipschitz optimization of (Hansen et al., 1992; Jones et al., 1993) and they led to significant relaxations of the conditions required for convergence, e.g. only the existence of a local *smoothness* around the optimum is required (Munos, 2011; Grill et al., 2015). More precisely, in the work of (Bull, 2011; Munos, 2011), specific conditions have been identified to derive a finite-time analysis of the algorithms. However, these guarantees do not hold when the unknown function is not assumed to be locally smooth around (one of) its optimum.

In the present work, we propose to explore concepts from ranking theory based on overlaying estimated level sets (Clémençon et al., 2010) in order to develop global optimization algorithms that do not rely on the smoothness of the function. The idea behind this approach is simple: even if the unknown function presents arbitrary large variations, most of the information required to identify its optimum may be contained in its induced ranking rule, i.e. how the level sets of the function are included one in another. To exploit this idea, we introduce a novel optimization scheme where the complexity of the function is characterized by the underlying pairwise ranking which it defines. Our contribution is twofold: first, we introduce two novel global optimization algorithms that learn the ranking rule induced by the unknown function with a sequential scheme, and second, we provide mathematical results in terms of statistical consistency and convergence to the optimum. Moreover the algorithms proposed lead to efficient implementation and they display good performance on the classical benchmarks for global optimization as shown at the end of the paper.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

The rest of the paper is organized as follows. In Section 2 we introduce the framework and give the main definitions. In Section 3, we introduce and analyze the RANKOPT algorithm which requires a prior information on the ranking structure underlying the unknown function. In Section 4, an adaptive version of the algorithm is presented. Companion results which establish the equivalence between learning algorithms and optimization procedures are discussed in Section 5 as they support implementation choices. The adaptive version of the algorithm is compared to other global optimization algorithms in Section 6. Proof sketches are postponed to the Appendix section and full proofs can be found in the supplementary material provided as a separate document.

2. Global optimization and ranking structure

2.1. Setup and notations

Setup. We consider the problem of sequentially maximizing an unknown real-valued function $f : \mathcal{X} \to \mathbb{R}$ which is assumed to admit at least one global maximum over the compact and convex set $\mathcal{X} \subset \mathbb{R}^d$. The objective is to identify some point

$$x^\star \in \arg\max_{x \in \mathcal{X}} f(x)$$

with a minimal amount of function evaluations. The setup we consider is the following: at each iteration $t = 1 \dots n-1$, an algorithm selects an evaluation point $X_{t+1} \in \mathcal{X}$ which depends on the previous evaluations $\{(X_i, f(X_i))\}_{i=1}^t$ and receives the evaluation of the unknown function $f(X_{t+1})$ at this point. After n iterations, the algorithm returns the argument of the highest value observed so far:

$$X_{\hat{\imath}_n} \quad \text{where} \quad \hat{\imath}_n \in \arg\max_{i=1,\dots,n} f(X_i).$$

The analysis provided in the paper considers that the number n of evaluation points is not fixed and it is assumed that function evaluations are noiseless.

Notations. For any $x = \{x_1 \dots x_d\} \in \mathbb{R}^d$, we define the standard ℓ_2-norm $\|x\|_2^2 = \sum_{i=1}^d x_i^2$, we denote by $\langle \cdot, \cdot \rangle$ the corresponding inner product and we denote by $B(x, r) = \{x' \in \mathbb{R}^d : \|x - x'\|_2 \leq r\}$ the ℓ_2-ball of radius $r \geq 0$ centered in x. For any set $\mathcal{X} \subset \mathbb{R}^d$, we define its inner-radius as $\mathrm{rad}(\mathcal{X}) = \max\{r > 0 : \exists x \in \mathcal{X} \text{ s.t. } B(x, r) \subseteq \mathcal{X}\}$, its diameter as $\mathrm{diam}(\mathcal{X}) = \max_{(x,x') \in \mathcal{X}^2} \|x - x'\|_2$ and we denote by $\mu(\mathcal{X})$ its volume where μ stands for the Lebesgue measure. Finally, we denote by $\mathcal{C}^0(\mathcal{X}, \mathbb{R})$ the set of continuous functions defined on $\mathcal{X}$ taking values in $\mathbb{R}$, we denote by $\mathcal{P}_N(\mathcal{X}, \mathbb{R})$ the set of (multivariate) polynomial functions of degree N defined on $\mathcal{X}$ and we denote by $\mathcal{U}(\mathcal{A})$ the uniform distribution over a bounded measurable domain $\mathcal{A}$.

2.2. The ranking structure of a real-valued function

In this section, we introduce the ranking structure as a complexity characterization for a general real-valued function to be optimized. First, we observe that real-valued functions induce an order relation over the input space $\mathcal{X}$, and the underlying ordering induces a ranking rule which records pairwise comparisons between evaluation points.

Definition 1. (INDUCED RANKING RULE) *The ranking rule $r_f : \mathcal{X} \times \mathcal{X} \to \{-1, 0, 1\}$ induced by a function $f : \mathcal{X} \to \mathbb{R}$ is defined by:*

$$r_f(x, x') = \begin{cases} 1 & \text{if } f(x) > f(x') \\ 0 & \text{if } f(x) = f(x') \\ -1 & \text{if } f(x) < f(x') \end{cases}$$

for all $(x, x') \in \mathcal{X}^2$.

The key argument of the paper is that the optimization of any weakly regular real-valued function only depends on the nested structure of its level sets. Hence there is an equivalence class of real-valued functions that share the same induced ranking rule as shown by the following proposition.

Proposition 1. (RANKING RULE EQUIVALENCE) *Let $h \in \mathcal{C}^0(\mathcal{X}, \mathbb{R})$ be any continuous function. Then, a function $f : \mathcal{X} \to \mathbb{R}$ shares the same induced ranking rule with h (i.e. $r_f = r_h$) if and only if there exists a strictly increasing (not necessary continuous) function $\psi : \mathbb{R} \to \mathbb{R}$ such that $h = \psi \circ f$.*

Proposition 1 states that even if the unknown function f admits non-continuous or large variations, up to a transformation ψ, there might exist a simpler function $h = \psi \circ f$ that shares the same induced ranking rule. Figure 1 gives an example of two functions that share the same ranking while they display highly different regularity properties. As a second example, we may consider the problem of maximizing the function $f(x) = 1 - 1/|\ln(x)|$ if $x \neq 0$ and 1 otherwise over $\mathcal{X} = [0, 1/2]$. The function f in this case is not *'smooth'* around its unique global maximizer $x^\star = 0$ but shares the same induced ranking rule with $h(x) = -x$ over $\mathcal{X}$.

We can now introduce a complexity characterization of real-valued functions of a set $\mathcal{X}$ through the complexity class of its induced ranking rule. We call this class a ranking structure.

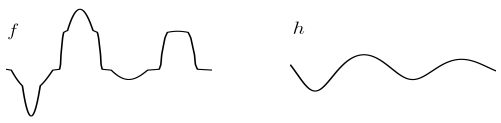


Figure 1. Two functions f and h that share the same ranking

Definition 2. (CONTINUOUS RANKING STRUCTURE AND CONTINUOUS RANKING RULES) *We say that a real-valued function f has a continuous ranking rule if $r_f \in \mathcal{R}_\infty$ where $\mathcal{R}_\infty = \{r_h : h \in \mathcal{C}^0(\mathcal{X}, \mathbb{R})\}$ denotes the set of continuous ranking rules (i.e. ranking rules induced by continuous functions).*

In the continuation of this definition, we further introduce two examples of more stringent ranking structures.

Definition 3. (POLYNOMIAL RANKING RULES) *The set of polynomial ranking rules of degree $N \geq 1$ is defined as*

$$\mathcal{R}_{\mathcal{P}(N)} = \{r_h : \ h \in \mathcal{P}_N(\mathcal{X}, \mathbb{R})\}.$$

We point out that even a polynomial function of degree N may admit a lower degree polynomial ranking rule. For example, consider the polynomial function $f(x) = (x^2 - 3x+1)^9$. Since $f(x) = \psi(x^2-3x)$ where $\psi : x \mapsto (x+1)^9$ is a strictly increasing function, the ranking rule induced by f is a polynomial ranking rule of degree 2.

The second class of ranking structures we introduce is a class of non-parametric rankings.

Definition 4. (CONVEX RANKING RULES) *The set of convex ranking rules of degree $N \geq 1$ is defined as*

$$\mathcal{R}_{\mathcal{C}(N)} = \{r \in \mathcal{R}_\infty \text{ such that for any } x' \in \mathcal{X}, \text{ the set } \{x \in \mathcal{X} : r(x, x') \geq 0\} \text{ is a union of } N \text{convex sets}\}.$$

It is easy to see that the ranking rule of a function f is a convex ranking rule of degree N if and only all the level sets of the function f are unions of at most N convex sets.

2.3. Identifiability and regularity

We now state two conditions that will be used in the theoretical analysis: the first condition is about the identifiability of the maximum of the function and the second is about the regularity of f around its maximum.

Condition 1. (IDENTIFIABILITY) *The maximum of a function $f : \mathcal{X} \to \mathbb{R}$ is said to be identifiable if for any $\varepsilon > 0$ arbitrarily small,*

$$\mu(\{x \in \mathcal{X} : f(x) \geq \max_{x \in \mathcal{X}} f(x) - \varepsilon\}) > 0.$$

Condition 1 prevents the function from having a jump on its maximum and will be useful to state asymptotic results of the type $f(X_{\hat{\imath}_n}) \to \max_{x \in \mathcal{X}} f(x)$ when $n \to +\infty$.

Condition 2. (REGULARITY OF THE LEVEL SETS) *A function $f : \mathcal{X} \to \mathbb{R}$ has (c_α, α)-regular level sets for some $c_\alpha > 0$, $\alpha \geq 0$ if:*

1. The global optimizer $x^\star \in \mathcal{X}$ is unique.

2. For any $y \in \mathrm{Im}(f)$, the iso-level set $f^{-1}(y) = \{x \in \mathcal{X} : f(x) = y\}$ satisfies

$$\max_{x \in f^{-1}(y)} \|x^\star - x\|_2 \leq c_\alpha \cdot \min_{x \in f^{-1}(y)} \|x^\star - x\|_2^{1/(1+\alpha)}.$$

Condition 2 guarantees that the points associated with high evaluations are close to the unique optimizer with respect to the Euclidean distance. This condition will be used to derive some finite-time bounds on the distance $\|x^\star - X_{\hat{\imath}_n}\|_2$ between the optimizer and its estimation. Note that for any iso-level set $f^{-1}(y)$ with finite distance to the optimum, the condition is satisfied with $\alpha = 0$ and $c_\alpha = \mathrm{diam}(\mathcal{X}) \,/\, \min_{x \in f^{-1}(y)} \|x^\star - x'\|_2$. Therefore, this condition concerns the behavior of the level sets when $\min_{x \in f^{-1}(y)} \|x^\star - x\|_2 \to 0$. As an example, the iso-level sets of three simple functions satisfying the condition with different values of α are shown in Figure 2.

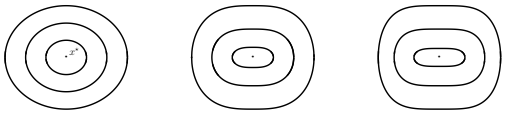


Figure 2. Illustration of the regularity of the level sets on two simple functions. *Left:* $f(x_1, x_2) = \exp(-x_1^2 - 2x_2^2)$ where $\alpha = 0$. *Middle:* $f(x) = -|x_1|^3 - 2x_2^2$ where $\alpha = 1/2$. *Right:* $f(x_1, x_2) = -x_1^4 - 2x_2^2$ where $\alpha = 1$.

3. Optimization with fixed ranking structure

In this section, we consider the problem of optimizing an unknown function f given the prior knowledge that its ranking r_f belongs to a given ranking structure $\mathcal{R} \subseteq \mathcal{R}_\infty$.

3.1. The RANKOPT algorithm

The input of **Algorithm 1** are a number n of iterations, the unknown function f, a compact and convex set $\mathcal{X} \subset \mathbb{R}^d$ and a ranking structure $\mathcal{R} \subseteq \mathcal{R}_\infty$. At each iteration $t < n$, a point X_{t+1} is uniformly sampled over $\mathcal{X}$ and the algorithm decides, whether or not, to evaluate the function at this point. The decision rule involves the active subset of $\mathcal{R}$ which contains the ranking rules that are consistent with the ranking rule induced by f over the points sampled so far. We thus set $\mathcal{R}_t = \{r \in \mathcal{R} : L_t(r) = 0\}$ where L_t is the empirical ranking loss:

$$L_t(r) = \frac{2}{t(t+1)} \sum_{1 \leq i < j \leq t} \mathbb{1}\{r_f(X_i, X_j) \neq r(X_i, X_j)\}.$$

Algorithm 1 RANKOPT$(n, f, \mathcal{X}, \mathcal{R})$

1. Initialization: Let $X_1 \sim \mathcal{U}(\mathcal{X})$
Evaluate $f(X_1)$, $t \leftarrow 1$, $\mathcal{R}_1 \leftarrow \mathcal{R}$, $\hat{\imath}_1 \leftarrow 1$
2. Iterations: Repeat while $t < n$
Let $X_{t+1} \sim \mathcal{U}(\mathcal{X})$
If there exists $r \in \mathcal{R}_t$ such that $r(X_{t+1}, X_{\hat{\imath}_t}) \geq 0$
Evaluate $f(X_{t+1})$, $t \leftarrow t+1$
$\mathcal{R}_t \leftarrow \{r \in \mathcal{R} : L_t(r) = 0\}$
$\hat{\imath}_t \in \arg\max_{i=1\ldots t} f(X_i)$
3. Output: Return $X_{\hat{\imath}_n}$

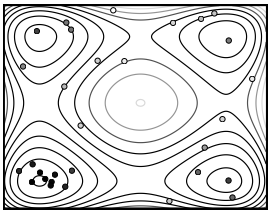
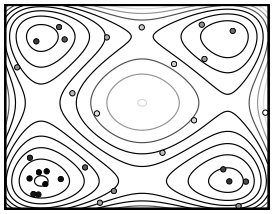



Figure 3. Two samples generated by the RANKOPT algorithm after $n = 30$ iterations with the polynomial ranking rules $\mathcal{R}_{\mathcal{P}(4)}$ on the Styblinski-Tang function defined in Section 6.

As a direct consequence of the definition of the active subset, if there does not exist any $r \in \mathcal{R}_t$ such that $r(X_{t+1}, X_{\hat{\imath}_n}) \geq 0$, it implies that $r_f(X_{t+1}, X_{\hat{\imath}_t}) = -1$ which means that $f(X_{t+1}) < f(X_{\hat{\imath}_t})$. Thus, the algorithm never evaluates the function at a point that will not return certainly an evaluation at least equal to the highest evaluation $f(X_{\hat{\imath}_t})$ observed so far.

Remark 1. (APPROXIMATION OF THE LEVEL SETS) Since r_f can be any ranking of $\mathcal{R}_t$, at each iteration, the sampling area $\mathcal{X}_t = \{x \in \mathcal{X} : \exists r \in \mathcal{R}_t \text{ s.t. } r(x, X_{\hat{\imath}_t}) \geq 0\}$ is the smallest set that contains certainly the level set $\{x \in \mathcal{X} : f(x) \geq f(X_{\hat{\imath}_t})\}$ of the best value observed so far.

Remark 2. (CONNECTION WITH ACTIVE LEARNING) The algorithm can be seen as an extension to ranking of the baseline *active learning* algorithm CAL (Cohn et al., 1994; Hanneke, 2011). However, in active learning we aim at estimating a classifier $h : \mathcal{X} \mapsto \{0, 1\}$ where the goal in global optimization is to estimate the winner of a tournament deriving from the ranking rule $r_f : \mathcal{X} \times \mathcal{X} \mapsto \{-1, 0, 1\}$ and not the ranking rule itself.

3.2. Convergence analysis

We state here some convergence properties of the RANKOPT algorithm. The results are stated in a probabilistic framework. The source of randomness comes from the algorithm itself (which generates uniform random variables) and not from the evaluations which are assumed noiseless. The next result will be important in order to formulate the consistency property of the algorithm.

Proposition 2. *Fix any $n \in \mathbb{N}^\star$, let $\mathcal{X} \subset \mathbb{R}^d$ be any compact and convex set and let $\mathcal{R} \subseteq \mathcal{R}_\infty$ be any set of ranking rules. Then, for any function $f : \mathcal{X} \to \mathbb{R}$ such that $r_f \in \mathcal{R}$ and any $y \in \mathbb{R}$, if $X_{\hat{\imath}_n}$ denotes the random output of* RANKOPT$(n, f, \mathcal{X}, \mathcal{R})$*, we have that,*

$$\mathbb{P}(f(X_{\hat{\imath}_n}) \geq y) \geq \mathbb{P}(\max_{i=1\ldots n} f(X'_i) \geq y),$$

where $\{X'_i\}_{i=1}^n$ are n independent random variables, uniformly distributed over $\mathcal{X}$.

Combining the previous proposition with the identifiability condition gives the following asymptotic result.

Corollary 1. (CONSISTENCY) *Using the same notations and assumptions as in Proposition 2 and if the maximum of the function f is identifiable (Condition 1), then,*

$$f(X_{\hat{\imath}_n}) \xrightarrow{\mathbb{P}} \max_{x \in \mathcal{X}} f(x).$$

We now provide our finite-sample loss bounds.

Theorem 1. (UPPER BOUND) *Under the same assumptions as in Proposition 2 and if the function f has (c_α, α)-regular level sets (Condition 2), then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\|x^\star - X_{\hat{\imath}_n}\|_2 \leq C_{\alpha,\mathcal{X}} \cdot \left(\frac{\ln(1/\delta)}{n}\right)^{\frac{1}{d(1+\alpha)^2}}$$

where $C_{\alpha,\mathcal{X}} = c_\alpha^{(2+\alpha)/(1+\alpha)} \operatorname{diam}(\mathcal{X})^{1/(1+\alpha)^2}$.

More surprisingly, a lower bound can be derived by making the link with the theoretical PURE ADAPTIVE SEARCH (Zabinsky & Smith, 1992) that uses the knowledge of the level sets of the unknown function.

Proposition 3. *Fix any $n \in \mathbb{N}^\star$ and let $\{X_i^\star\}_{i=1}^n$ be a sequence distributed as the Markov process defined by*

$$\begin{cases} X_1^\star \sim \mathcal{U}(\mathcal{X}) \\ X_{t+1}^\star | \; X_t^\star \sim \mathcal{U}(\mathcal{X}_t^\star) \qquad \forall t \in \{1 \ldots n-1\} \end{cases}$$

where $\mathcal{X}_t^\star = \{x \in \mathcal{X} : f(x) \geq f(X_t^\star)\}$. Then, using the same notations and assumptions as in Proposition 2, for any $y \in \mathbb{R}$, we have that,

$$\mathbb{P}(f(X_{\hat{\imath}_n}) \geq y) \leq \mathbb{P}(f(X_n^\star) \geq y).$$

We are now ready to establish our second loss bound by combining Proposition 3 with the level set assumption.

Theorem 2. (LOWER BOUND) *Under the same assumptions as in Proposition 2 and if the function f has (c_α, α)-regular level sets (Condition 2), then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$C_{\alpha,\mathcal{X}} \cdot e^{-\frac{(1+\alpha)^2}{d}\left(n + \sqrt{2n\ln(1/\delta)} + \ln(1/\delta)\right)} \leq \|x^\star - X_{\hat{\imath}_n}\|_2,$$

where $C_{\alpha,\mathcal{X}} = c_\alpha^{-(1+\alpha)(2+\alpha)} \operatorname{rad}(\mathcal{X})^{(1+\alpha)^2}$.

Remark 3. (ON THE PERFORMANCE CRITERION) The level set assumption, which is used in Theorem 1 and Theorem 2, is invariant to any strictly increasing composition ψ (i.e. if f has (c_α, α)-regular level sets so has $\psi \circ f$). It implies that the bounds on the distance $\|x^\star - X_{\hat{\imath}_n}\|_2$ between the exact solution and its approximation hold independently of the smoothness of the function.

Remark 4. (RELATED WORK) To the best of our knowledge, this is the first analysis of an optimization algorithm

which uses the ranking rule induced by the unknown function. For a different approach to global optimization, we refer to the work of (Munos, 2011) where the function is assumed to be locally smooth around (one of) its optimum. In the latter work, finite-time bounds on the difference $\max_{x\in\mathcal{X}} f(x) - f(X_{\hat{\imath}_n})$ are obtained when the smoothness is known (DOO) and the smoothness is unknown (SOO).

4. Adaptive algorithm and stopping time analysis

4.1. The AdaRankOpt algorithm

The AdaRankOpt algorithm (**Algorithm 2**) is an extension of the RankOpt algorithm which involves model selection following the principle of Structural Risk Minimization. We consider a parameter $p \in (0,1)$ and a nested sequence of ranking structures $\{\mathcal{R}_N\}_{N\in\mathbb{N}^\star}$ satisfying:

$$\mathcal{R}_1 \subset \mathcal{R}_2 \subset \cdots \subset \mathcal{R}_\infty. \tag{1}$$

The algorithm is initialized by considering the smallest ranking structure $\mathcal{R}_1$ of the sequence. At each iteration $t < n$, a Bernoulli random variable B_{t+1} of parameter p is sampled. If $B_{t+1} = 1$, the algorithm explores the space by evaluating the function at a point uniformly sampled over $\mathcal{X}$. If $B_{t+1} = 0$, the algorithm exploits the previous evaluations by making an iteration of the RankOpt algorithm with the smallest ranking structure $\mathcal{R}_{N_t}$ of the sequence that probably contains the true ranking r_f. Once a new evaluation has been made, the index N_{t+1} is updated. The parameter p drives the trade-off between the exploitation phase and the exploration phase which prevents the algorithm from getting stuck in a local maximum.

Remark 5. (Nested sequences) Condition (1) is crucial for practical reasons discussed in Section 5. We point out that both the sequence of polynomial ranking rules $\{\mathcal{R}_{\mathcal{P}(N)}\}_{N\in\mathbb{N}^\star}$ and the sequence of convex ranking rules $\{\mathcal{R}_{\mathcal{C}(N)}\}_{N\in\mathbb{N}^\star}$ defined in Section 2 satisfy this condition.

Algorithm 2 AdaRankOpt$(n, f, \mathcal{X}, p, \{\mathcal{R}_N\}_{N\in\mathbb{N}^\star})$

1. Initialization: Let $X_1 \sim \mathcal{U}(\mathcal{X})$
Evaluate $f(X_1)$, $t \leftarrow 1$, $\mathcal{R} \leftarrow \mathcal{R}_1$, $\hat{\imath}_1 \leftarrow 1$
2. Iterations: Repeat while $t < n$
Let $B_{t+1} \sim \mathcal{B}(p)$
If $B_{t+1} = 1$ {**Exploration**}
  Let $X_{t+1} \sim \mathcal{U}(\mathcal{X})$
If $B_{t+1} = 0$ {**Exploitation**}
  Let $X_{t+1} \sim \mathcal{U}(\{x \in \mathcal{X} : \exists r \in \mathcal{R} \text{ s.t. } r(x, X_{\hat{\imath}_t}) \geq 0\})$
Evaluate $f(X_{t+1})$, $t \leftarrow t+1$
$\hat{\imath}_t \in \arg\max_{i=1\ldots t} f(X_i)$
$N_t \leftarrow \min\{N \in \mathbb{N}^\star : \min_{r\in\mathcal{R}_N} L_t(r) = 0\}$ {**Update**}
$\mathcal{R} \leftarrow \{r \in \mathcal{R}_{N_t} : L_t(r) = 0\}$
3. Output: Return $X_{\hat{\imath}_n}$

4.2. Theoretical properties of AdaRankOpt

We start by casting the consistency result for the AdaRankOpt algorithm.

Proposition 4. (Consistency) *Fix any $p \in (0,1)$ and any sequence of ranking structures $\{\mathcal{R}_N\}_{N\in\mathbb{N}^\star}$ satisfying (1). Then, if the function f has an identifiable maximum (Condition 1) and $X_{\hat{\imath}_n}$ denotes the random output of* AdaRankOpt$(n, f, \mathcal{X}, p, \{\mathcal{R}_N\}_{N\in\mathbb{N}^\star})$*, we have that,*

$$f(X_{\hat{\imath}_n}) \xrightarrow{\mathbb{P}} \max_{x\in\mathcal{X}} f(x).$$

Proposition 4 reveals that even if the algorithm is poorly tuned, it will end up finding the true maximum of any function with an identifiable maximum.

We now investigate the number of iterations required to identify a ranking structure that contains the *true* ranking rule.

Definition 5. (Stopping Time) *Let $N^\star = \min\{N \in \mathbb{N}^\star : r_f \in \mathcal{R}_N\}$ be the index of the smallest ranking structure that contains the true ranking rule and let $\{N_t\}_{t\in\mathbb{N}^\star}$ be the sequence of random variables defined in the* AdaRankOpt *algorithm. Define the stopping time:*

$$\tau_{N^\star} = \min\{t \in \mathbb{N}^\star : N_t = N^\star\}$$

which corresponds to the number of iterations required to identify $N^\star$.

In order to bound $\tau_{N^\star}$, we need to control the complexity of the sequence of ranking structures. Let us denote by $L(r) = \mathbb{P}(r_f(X, X') \neq r(X, X'))$ the true ranking loss where (X, X') is a couple of independent random variables uniformly distributed over $\mathcal{X}$ and define the Rademacher average of a ranking structure $\mathcal{R}$ as

$$R_n = \sup_{r\in\mathcal{R}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \epsilon_i \mathbb{1}_{[r_f(X_i, X_{\lfloor n/2 \rfloor + i}) \neq r(X_i, X_{\lfloor n/2 \rfloor + i})]} \right|$$

where $\{X_i\}_{i=1}^n$ are n i.i.d. copies of $X \sim \mathcal{U}(\mathcal{X})$ and $\epsilon_1 \ldots \epsilon_{\lfloor n/2 \rfloor}$ are $\lfloor n/2 \rfloor$ independent Rademacher random variables (*i.e.* random symmetric sign variables), independent of $\{X_i\}_{i=1}^n$.

Proposition 5. (Stopping Time Upper Bound) *Assume that the index $N^\star > 1$ is finite, and that $\inf_{r\in\mathcal{R}_{N^\star-1}} L(r) > 0$, and that there exists $V > 0$ such that the Rademacher complexity of $\mathcal{R}_{N^\star-1}$ satisfies $\mathbb{E}[R_n] \leq \sqrt{V/n}$ for all $n \in \mathbb{N}^\star$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\tau_{N^\star} \leq \frac{10}{p} \cdot \left(\frac{V + \ln(2/\delta)}{\inf_{r\in\mathcal{R}_{N^\star-1}} L(r)^2} \right).$$

In the situation described above, one can recover an upper bound similar to the one of Theorem 1 by combining the previous result with the analysis of the RANKOPT algorithm where the structure $\mathcal{R}_{N^\star}$ is assumed to be known.

Theorem 3. (UPPER BOUND) *Consider the same assumptions as in Proposition 5 and assume that the function f has (c_α, α)-regular level sets (Condition 2). Then, if $X_{\hat{\imath}_n}$ denotes the random output of* ADARANKOPT$(n, f, \mathcal{X}, p, \{\mathcal{R}_N\}_{N\in\mathbb{N}^\star})$, *for any $\delta \in (0,1)$ and any $n > n_\delta$, with probability at least $1-\delta$,*

$$\|x^\star - X_{\hat{\imath}_n}\|_2 \leq C_{\alpha,\mathcal{X}} \cdot \left(\frac{\ln(2/\delta)}{n - n_\delta}\right)^{\frac{1}{d(1+\alpha)^2}}$$

where $C_{\alpha,\mathcal{X}}$ is the same constant as in Theorem 1 and $n_\delta = \lfloor 10(V + \ln(4/\delta))/(p \cdot \inf_{r\in\mathcal{R}_{N^\star-1}} L(r)^2)\rfloor$.

Remark 6. (COMPLEXITY ASSUMPTION) We refer here to (Clémençon, 2011) and we point out that standard VC-type arguments can be used in order to bound $\mathbb{E}[R_n]$. If the set of functions $\mathcal{F} = \{(x,x') \in \mathcal{X}^2 \mapsto \mathbb{1}\{r_f(x,x') \neq r(x,x')\} : r \in \mathcal{R}\}$ is a VC major class with finite VC dimension V, then $\mathbb{E}[R_n] \leq c\sqrt{V/n}$ for a universal constant $c > 0$. This covers the case of polynomial ranking rules.

5. Computational aspects

We discuss here some technical aspects involved in the practical implementation of the ADARANKOPT algorithm.

5.1. General ranking structures

Fix any nested sequence of ranking structures $\{\mathcal{R}_N\}_{N\in\mathbb{N}^\star}$ and any sample $\{(X_i, f(X_i))\}_{i=1}^n$. We address the questions of (i) sampling X_{n+1} uniformly over the non-trivial subset $\mathcal{X}_n = \{x \in \mathcal{X} : \exists r \in \mathcal{R}_{N_n} \text{ s.t. } L_n(r) = 0 \text{ and } r(x, X_{\hat{\imath}_n}) \geq 0\}$ and (ii) updating the index N_{n+1} once $f(X_{n+1})$ has been evaluated. We start to show that both these steps can be done by testing if

$$\min_{r\in\mathcal{R}_N} L_{n+1}(r) = 0 \tag{2}$$

holds true for a given $N \in \mathbb{N}^\star$ where the empirical ranking loss is taken over a set of $n+1$ samples.

(i) Sampling $X \sim \mathcal{U}(\mathcal{X})$ until $X \in \mathcal{X}_n$ allows to sample uniformly over $\mathcal{X}_n$. Using the definition of the subset, we know that $X \in \mathcal{X}_n$ if there exists a ranking $r \in \mathcal{R}_{N_n} \cap \{r : L_n(r) = 0\}$ such that $r(X, X_{\hat{\imath}_n}) = 0$ *or* 1. Rewriting the previous statement in terms of minimal error gives that $X \in \mathcal{X}_n$ if:

- either $\min_{r\in\mathcal{R}_{N_n}} L_{n+1}(r) = 0$ where L_{n+1} is taken over the sample $\{(X_i, f(X_i))\}_{i=1}^n \cup (X, f(X_{\hat{\imath}_n}))$,
- or $\min_{r\in\mathcal{R}_{N_n}} L_{n+1}(r) = 0$ where L_{n+1} is taken over the sample $\{(X_i, f(X_i))\}_{i=1}^n \cup (X, f(X_{\hat{\imath}_n}) + c)$ where $c > 0$ is any strictly positive constant.

(ii) Assume now that $f(X_{n+1})$ has been evaluated. Since $\{\mathcal{R}_N\}_{N\in\mathbb{N}^\star}$ forms a nested sequence, we have that $N_{n+1} = N_n + \min\{i \in \mathbb{N}^\star : \min_{r\in\mathcal{R}_{N_n+i}} L_{n+1}(r) = 0\}$ where the empirical loss is taken over $\{(X_i, f(X_i))\}_{i=1}^{n+1}$. Therefore, N_{n+1} can be updated by sequentially testing if $\min_{r\in\mathcal{R}_{N_n+i}} L_{n+1}(r) = 0$ for $i = 0, 1, 2 \ldots$

As mentioned earlier, both the previous steps can be done using a generic procedure that tests if (2) holds true. We now provide some equivalences that can be used to design this procedure for the ranking structures introduced in Section 2. For simplicity, we assume that all the evaluations of the sample are distinct:

$$f(X_{(1)}) < f(X_{(2)}) < \ldots < f(X_{(n+1)}). \tag{3}$$

where $(1) \ldots (n+1)$ denote the indexes of the corresponding reordering.

5.2. Polynomial ranking rules

Consider the sequence of polynomial ranking rules $\{\mathcal{R}_{\mathcal{P}(N)}\}_{N\in\mathbb{N}^\star}$ and let $\phi_N : \mathbb{R}^d \to \mathbb{R}^{\dim(\phi_N)}$ be the function that maps any point of $\mathbb{R}^d$ into the polynomial feature space of degree N where $\dim(\phi_N) = \binom{N+d}{d} - 1$. For example, $\phi_2(x_1, x_2) = \{x_1, x_2, x_1x_2, x_1^2, x_2^2\}$. We start by making the link with linear separability in the polynomial feature space.

Proposition 6. (LINEAR SEPARABILITY) *Fix any $N \in \mathbb{N}^\star$ and assume that all the evaluations are distinct (3). Then, (2) holds true if and only if there exists an axis $\omega \in \mathbb{R}^{\dim(\phi_N)}$ such that,*

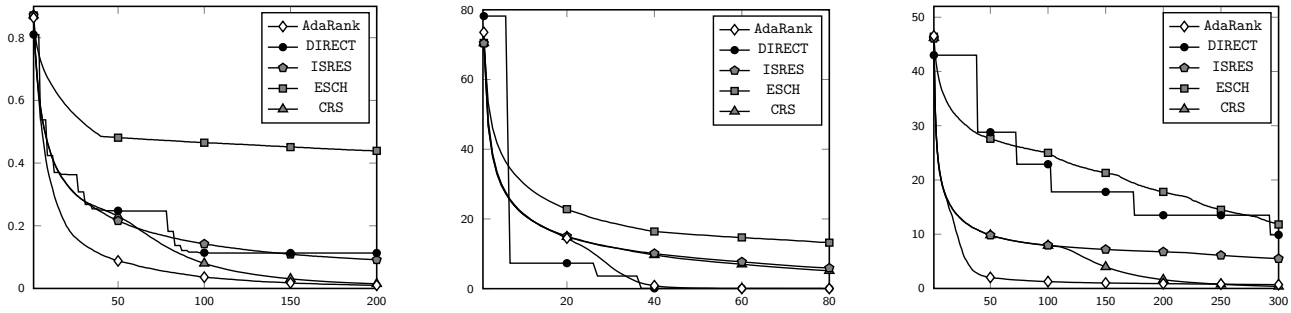
$$\langle \omega, \phi_N(X_{(i+1)}) - \phi_N(X_{(i)})\rangle > 0, \ \forall i \in \{1 \ldots n\}.$$

Interestingly, testing the linear separability of a sample is equivalent to testing the emptiness of a sample-dependent polyhedron.

Corollary 2. *Let M_N be the $(\dim(\phi_N), n)$-matrix where its i-th column is equal to $(\phi_N(X_{(i+1)}) - \phi_N(X_{(i)}))^T$ and let the operator $\succeq$ stands for the inequality $\geq$ componentwise (i.e. $x \succeq x' \Leftrightarrow x_i \geq x'_i \ \forall i \in \{1 \ldots d\}$). Then, under the same assumptions as in Proposition 6, (2) holds true if and only if the following polyhedron is empty:*

$$\{\lambda \in \mathbb{R}^n : \ \mathrm{M}_N\lambda^T = \vec{0}, \ \langle \vec{1}, \lambda\rangle = 1, \ \lambda \succeq \vec{0}\} = \emptyset\,.$$

Remark. 7 (ALGORITHMIC ASPECTS) The problem of testing the emptiness of a polyhedron can be seen as the problem of finding a feasible point of a linear program. We refer to Chapter 11.4 in (Boyd & Vandenberghe, 2004) where algorithmic solutions are discussed.



(a) **(b)** **(c)**

Figure 4. Empirical estimation of $\mathbb{E}\left[\max_{x\in\mathcal{X}} f(x) - f(X_{i_t})\right]$ in terms of iteration $t = 1 \dots n$ where the expectation was obtained by running a 1000 times each algorithm.

5.3. Convex ranking rules

Consider the sequence of convex ranking rules $\{\mathcal{R}_{\mathcal{C}(N)}\}_{N\in\mathbb{N}^\star}$. Following the steps of (Clémençon & Vayatis, 2010) leads to the next equivalence.

Proposition 7. (OVERLAYING CLASSIFIERS) *Fix any $N \in \mathbb{N}^\star$ and let $\mathcal{X} = [a, b]$. Then, (2) holds true if and only if there a exists a nested sequence $h_1 \geq h_2 \geq \dots \geq h_{n+1}$ of $n+1$ classifiers of the form $h_i(x) = \sum_{k=1}^{N} \mathbb{1}\{l_{i,k} \leq x \leq u_{i,k}\}$ satisfying $h_i(X_{(j)}) = \mathbb{1}\{(j) \geq i\}$, $\forall (i, j) \in \{1 \dots n+1\}^2$.*

The problem of overlaying classifiers admits a tractable solution when $d = 1$. In the specific case where $N = 1$ and $d \in \mathbb{N}^\star$, the problem of testing the existence of nested convex classifiers is equivalent to the problem of testing the emptiness of a cascade of polyhedrons.

Proposition 8. *Fix any $d \in \mathbb{N}^\star$, let $N = 1$ and assume that all the evaluations are distinct. Then, (2) holds true if and only if for each $k \in \{1 \dots n\}$ the polyhedron defined by:*

$$\{\lambda \in \mathbb{R}^k : \mathrm{M}_k\lambda = X^T_{(n+1-k)},\ \langle\vec{1}, \lambda\rangle = 1,\ \lambda \succeq \vec{0}\}$$

where M_k is the (d, k)-matrix where its i-th column is equal to $X^T_{(n+2-i)}$, is empty.

6. Experiments

We now compare the empirical performances of the ADARANKOPT algorithm with four global optimization algorithms taken from the NLOpt library (Johnson, 2014):

CRS (Kaelo & Ali, 2006) is a controlled random search with local mutations. It starts with a random population and randomly evolve these points by an heuristic rule.

DIRECT (Jones et al., 1993) is a Lipschitz optimization algorithm where the Lipschitz constant is unknown. It uses a deterministic splitting technique of the search space.

ESCH (Santos et al., 2010) and ISRES (Runarsson & Yao, 2000) are two evolutionary algorithms. The evolution strategies are based on a combination of mutation rules and differential variations.

The tuning parameters were set to default and the parameter p was set to $1/4$ for the convex ranking rules and to $1/10$ for the polynomial ranking rules. The algorithms were compared on three synthetic problems:

(a) The task consists in maximizing the function $f(x) = \mathbb{1}\{x \leq x^\star\}(|\cos(50(x - x^\star))|^{3/2} - 15|x - x^\star|^{1/2})/10 - \mathbb{1}\{x > x^\star\}(|x|^{1/2} + 0.05|\sin(50x)|^{3/2})$ over $\mathcal{X} = [0, 1]$ where $x^\star = 0.499$. The function f has 17 local maxima and presents a discontinuity around its unique optimizer $x^\star$. The horizon n was set to 200 evaluations and the convex ranking rules were used.

(b) The task consists in minimizing a perturbed version of the Styblinski-Tang function $f(x) = \sum_{i=1}^{2}(x_i^4 - 16x_i^2 + 5x_i)/2 + \cos(x_1 + x_2)$ over $\mathcal{X} = [-5, 5]^2$. The level sets of the Styblinski- Tang function are displayed on Figure 3 and the function has 4 local minima. The polynomial ranking rules were used and the horizon n was set to 80 evaluations.

(c) The task consists in maximizing the function $f(x) = 1 - |\sum_{i=1}^{10}(x_i - 4.5)/10|^{5/2}$ over $\mathcal{X} = [-5, 5]^{10}$. The hyperplane $\{x \in \mathbb{R}^{10} : \sum_{i=1}^{10} x_i = 45\}$ maximizes the function. The horizon n was set to 300 evaluations and the polynomial ranking rules were used.

The results are shown in Figure 5.1. We remark that the ADARANKOPT converges fast and avoids falling in local maxima, as opposed to most of its competitors.

7. Conclusion

We have provided a global optimization strategy based on a sequential estimation of the ranking of the unknown function. We introduced two algorithms: RANKOPT which requires a prior knowledge of the ranking rule of the unknown function and its adaptive version ADARANKOPT which performs model selection. A theoretical analysis is provided and the adaptive algorithm is shown to be empirically competitive with the state-of-the-art methods on representative synthetic examples.

Appendix - Sketch of Proofs

Full proofs can be found in the Supplementary Material.

Proof of Proposition 1. The second inclusion ($\Leftarrow$) is a direct consequence of the definition of the ranking rules. To state the first inclusion ($\Rightarrow$), we introduce the function $M(x) = \mu(\{x' \in \mathcal{X} : r_f(x, x') = -1\})$ and we show that there exists two strictly increasing functions ψ and ψ' such that $f = \psi \circ M$ and $h = \psi' \circ M$. We finally get that $h = (\psi' \circ \psi^{-1}) \circ f$ where $\psi' \circ \psi^{-1}$ is strictly increasing.

Proof of Proposition 2. The result is obtained by induction. Since $X_1 \sim \mathcal{U}(\mathcal{X})$, the result trivially holds for $n = 1$. Assume that the statement holds for a given $n \in \mathbb{N}^\star$, fix any $y \in \mathbb{R}$ and let $\mathcal{X}_y = \{x \in \mathcal{X} : f(x) \geq y\}$. First step: using the fact that $\{x \in \mathcal{X} : f(x) \geq f(X_{\hat{\imath}_n})\} \subseteq \mathcal{X}_n \subseteq \mathcal{X}$ we show that $\mathbb{P}(f(X_{\hat{\imath}_{n+1}}) \geq y) \geq \mathbb{P}(f(X_{\hat{\imath}_n}) \geq y) + \mathbb{P}(f(X_{\hat{\imath}_n}) < y)\mu(\mathcal{X}_y)/\mu(\mathcal{X})$. Second step: plugging the induction assumption in the last equation and using the fact that $\mathbb{P}(f(X'_{n+1}) \geq y) = \mu(\mathcal{X}_y)/\mu(\mathcal{X})$ gives the result.

Proof of Corollary 1. Fix any $\varepsilon > 0$ and let $\mathcal{X}_\varepsilon = \{x \in \mathcal{X} : f(x) \geq \max_{x \in \mathcal{X}} f(x) - \varepsilon\}$ be the corresponding level set. Applying Proposition 2 leads to $\mathbb{P}(f(X_{\hat{\imath}_n}) < \max_{x \in \mathcal{X}} f(x) - \varepsilon) \leq (1 - \mu(\mathcal{X}_\varepsilon)/\mu(\mathcal{X}))^n \underset{n \to \infty}{\longrightarrow} 0$.

Proof of Theorem 1. Fix any $\delta \in (0, 1)$ and let $r_{\delta,n}$ be the upper bound of the theorem. First step: using Proposition 3 and the level set assumption, we show that $\mathbb{P}(\|X_{\hat{\imath}_n} - x^\star\|_2 \leq r_{\delta,n}) \geq \mathbb{P}(\cup_{i=1}^n \{X'_i \in B(x^\star, \operatorname{diam}(\mathcal{X})\,(\ln(1/\delta)/n)^{1/d})\})$ where $\{X'_i\}_{i=1}^n$ are n i.i.d. copies of $X' \sim \mathcal{U}(\mathcal{X})$. Second step: we show that for any $r \leq \operatorname{diam}(\mathcal{X})$ we have that $\mu(B(x^\star, r))/\mu(\mathcal{X}) \geq (r/\operatorname{diam}(\mathcal{X}))^d$. Third step: using independence and the fact that $1 - x \leq e^{-x}$, we finally get that $\mathbb{P}(\|X_{\hat{\imath}_n} - x^\star\|_2 \leq r_{\delta,n}) \geq 1 - (1 - \ln(1/\delta)/n)^n \geq 1 - \delta$.

Proof of Proposition 3. We use again an induction argument. Assume that the result holds for a given $n \in \mathbb{N}^\star$ and fix any $y \in \mathbb{R}$. First step: using the fact that $\mathcal{X}_{f(X_{\hat{\imath}_n})} = \{x \in \mathcal{X} : f(x) \geq f(X_{\hat{\imath}_n})\} \subseteq \mathcal{X}_n$ we show that $P(f(X_{\hat{\imath}_{n+1}}) \geq y) \leq \mathbb{E}\left[\min(1, \mu(\mathcal{X}_y)/\mu(\mathcal{X}_{f(X_{\hat{\imath}_n})}))\right]$. Second step: using the induction assumption we show that $\mathbb{E}\left[\min(1, \mu(\mathcal{X}_y)/\mu(\mathcal{X}_{f(X_{\hat{\imath}_n})}))\right] \leq \mathbb{E}\left[\min(1, \mu(\mathcal{X}_y)/\mu(\mathcal{X}_{f(X^\star_n)}))\right] = \mathbb{P}(f(X^\star_{n+1}) \geq y)$.

Proof of Theorem 2. Fix any $\delta \in (0, 1)$ and let $r_{\delta,n}$ be the lower bound of the corollary. First step: using Theorem 3 and the level set assumption we show that $\mathbb{P}(\|X_{\hat{\imath}_n} - x^\star\|_2 \leq r_{\delta,n}) \leq \mathbb{P}(\mu(\mathcal{X}^\star_n)/\mu(\mathcal{X}) \leq \delta \exp\left(-n - \sqrt{2n\ln(1/\delta)}\right))$ where $\mathcal{X}^\star_n = \{x \in \mathcal{X} : f(x) \geq f(X^\star_n)\}$. Second step: we show that $\forall u \in (0, 1)$, $\mathbb{P}(\mu(\mathcal{X}^\star_n)/\mu(\mathcal{X}) \leq u) \leq \mathbb{P}(\prod_{i=1}^n U_i \leq u)$ where $\{U_i\}_{i=1}^n$ are n i.i.d. copies of $U \sim \mathcal{U}([0, 1])$. Third step: using concentration inequalities for sub-gamma random variables gives that $\mathbb{P}(\prod_{i=1}^n U_i \leq \delta \exp\left(-n - \sqrt{2n\ln(1/\delta)}\right)) < \delta$.

Proof of Proposition 4. Fix any $\varepsilon > 0$ and let $\mathcal{X}_\varepsilon = \{x \in \mathcal{X} : f(x) \geq \max_{x \in \mathcal{X}} f(x) - \varepsilon\}$ be the corresponding level set. Using the fact that $\mathbb{P}(X_i \in \mathcal{X}_\varepsilon) \geq p \cdot \mu(\mathcal{X}_\varepsilon)/\mu(\mathcal{X})$ for any $i \in \mathbb{N}^\star$, we show by induction that $\mathbb{P}(f(X_{\hat{\imath}_n}) < \max_{x \in \mathcal{X}} f(x) - \varepsilon) \leq (1 - p \cdot \mu(\mathcal{X}_\varepsilon)/\mu(\mathcal{X}))^n \underset{n \to \infty}{\longrightarrow} 0$.

Proof of Proposition 5. Fix any $\delta \in (0, 1)$ and let n_δ be the integer part of the upper bound of the proposition. First step: since we have a nested sequence of sets of ranking rules, $\mathbb{P}(\tau \leq n_\delta) = \mathbb{P}(\min_{r \in \mathcal{R}_{N^\star - 1}} L_{n_\delta}(r) > 0)$. Second step: using Hoeffding's inequality gives a lower bound on the number of i.i.d. samples collected: $\mathbb{P}(\sum_{i=1}^{n_\delta} B_i \geq \lfloor p \cdot n_\delta - \sqrt{n_\delta \log(2/\delta)/2} \rfloor = n'_\delta) \geq 1 - \delta/2$. Third step: applying concentration results of ranking rules over the n'_δ i.i.d. samples gives that $\mathbb{P}(\min_{r \in \mathcal{R}_{N^\star - 1}} L_{n'_\delta}(r) \geq \min_{r \in \mathcal{R}_{N^\star - 1}} L(r) - 2\sqrt{V/n'_\delta} - 2\sqrt{\ln(2/\delta)/(n'_\delta - 1)} > 0) \geq 1 - \delta/2$.

Proof of Theorem 3. Fix any $\delta \in (0, 1)$. We know that after n_δ iterations the true ranking structure $\mathcal{R}_{N^\star}$ is identified with probability at least $1 - \delta/2$ (Proposition 5). Once the structure $\mathcal{R}_{N^\star}$ is identified, one can use a similar technique as the one used in Theorem 1 to get an upper bound with probability at least $1 - \delta/2$ thanks to the $n - n_\delta$ samples.

Proof of Proposition 6. The proof is a consequence of the definition of polynomial ranking rules: if $r \in \mathcal{R}_{\mathcal{P}(N)}$ then there exists $\omega_r \in \mathbb{R}^{\dim(\phi_N)}$ and $c_r \in \mathbb{R}$ such that $r(x, x') = \operatorname{sgn}(\langle \omega_r, \phi_N(x) \rangle + c_r - \langle \omega_r, \phi_N(x') \rangle - c_r) = \operatorname{sgn}(\langle \omega_r, \phi_N(x) - \phi_N(x') \rangle)$. Noticing that $r_f(X_{(i+1)}, X_{(i)}) = 1$ for all $i \in \{1 \dots n\}$ gives the result.

Proof of Corollary 2. Let $X'_i = \phi_N(X_{(i+1)}) - \phi_N(X_{(i)})$, $\forall i \in \{1 \dots n\}$ and let $\mathrm{CH}\{X'_i\}_{i=1}^n$ be the convex hull of $\{X'_i\}_{i=1}^n$. First step: we show the following equivalence: $\exists \omega \in \mathbb{R}^{\dim(\phi_N)}$ such that $\forall i \in \{1 \dots n\}$, $\langle \omega, X'_i \rangle > 0 \Leftrightarrow \vec{0} \notin \mathrm{CH}\{X'_i\}_{i=1}^n$. Second step: using the definition of convex hull we get the second equivalence: $\vec{0} \in \mathrm{CH}\{X'_i\}_{i=1}^n \Leftrightarrow \exists (\lambda_1, \dots \lambda_n) \in \mathbb{R}^n$ s.t. $\vec{0} = \sum_{i=1}^n \lambda_i X'_i$, $\sum_{i=1}^n \lambda_i = 1$ and $\lambda_i \geq 0, \forall i \in \{1 \dots n\}$.

Proof of Proposition 7. The first inclusion ($\Rightarrow$) is a direct consequence of the definition of convex ranking rules. Assume now that there exists a nested sequence of classifiers $h_1 \geq \dots \geq h_{n+1}$ satisfying the conditions. To state the second inclusion ($\Leftarrow$) we build a continuous approximation of the step function $f(x) = \sum_{i=1}^n h_i(x)$ that perfectly ranks the sample and induces a convex ranking.

Proof of Proposition 8. First step: using the definition of convex hulls, we show that each polyhedron of the cascade is empty iff $\mathrm{CH}\{X_{(i)}\}_{i=n}^{n+1} \subset \mathrm{CH}\{X'_{(i)}\}_{i=n-1}^{n+1} \subset \dots \subset \mathrm{CH}\{X'_{(i)}\}_{i=1}^{n+1}$. Second step: we build a continuous approximation of the function $f(x) = \sum_{k=1}^n \mathbb{1}\left\{x \in \mathrm{CH}\{X_{(i)}\}_{i=k}^{n+1}\right\}$ which induces a convex ranking rule and perfectly ranks the sample.

References

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Bull, Adam D. Convergence rates of efficient global optimization algorithms. *The Journal of Machine Learning Research*, 12:2879–2904, 2011.

Clémençon, Stéphan. On U-processes and clustering performance. In *Advances in Neural Information Processing Systems*, pp. 37–45, 2011.

Clémençon, Stéphan and Vayatis, Nicolas. Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648, 2010.

Clémençon, Stéphan, Lugosi, Gabor, and Vayatis, Nicolas. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pp. 844–874, 2010.

Cohn, David, Atlas, Les, and Ladner, Richard. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

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

Hanneke, Steve. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Hansen, Pierre, Jaumard, Brigitte, and Lu, Shi-Hui. Global optimization of univariate lipschitz functions: I. survey and properties. *Mathematical programming*, 55(1-3): 251–272, 1992.

Johnson, Steven G. The NLopt nonlinear-optimization package, 2014. http://ab-initio.mit.edu/nlopt.

Jones, Donald R, Perttunen, Cary D, and Stuckman, Bruce E. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

Jones, Donald R, Schonlau, Matthias, and Welch, William J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13 (4):455–492, 1998.

Kaelo, P and Ali, MM. Some variants of the controlled random search algorithm for global optimization. *Journal of optimization theory and applications*, 130(2):253–264, 2006.

Munos, Rémi. Optimistic optimization of deterministic functions without the knowledge of its smoothness. In *Advances in neural information processing systems*, 2011.

Pintér, János D. Global optimization in action. *Scientific American*, 264:54–63, 1991.

Rios, Luis Miguel and Sahinidis, Nikolaos V. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.

Runarsson, Thomas P and Yao, Xin. Stochastic ranking for constrained evolutionary optimization. *Evolutionary Computation, IEEE Transactions on*, 4(3):284–294, 2000.

Santos, Carlos Henrique da Silva, Goncalves, Marcos Sergio, and Hernandez-Figueroa, Hugo Enrique. Designing novel photonic devices by bio-inspired computing. *Photonics Technology Letters, IEEE*, 22(15):1177–1179, 2010.

Sergeyev, Yaroslav D, Strongin, Roman G, and Lera, Daniela. *Introduction to global optimization exploiting space-filling curves*. Springer Science & Business Media, 2013.

Zabinsky, Zelda B. *Stochastic adaptive search for global optimization*, volume 72. Springer Science & Business Media, 2013.

Zabinsky, Zelda B and Smith, Robert L. Pure adaptive search in global optimization. *Mathematical Programming*, 53(1-3):323–338, 1992.