Parallel Optimistic Optimization

Jean-Bastien Grill @ Inria, Michaël Valko @ Inria, and Munos @ Google

**Algorithm: Poo**

Parameters: $K$, $P = \{P_h\}$

Optional parameters: $\rho_{max}$, $\rho_{min}$

Initialization:

$D_{max} = \lim K/\ln(1/\rho_{min})$

$n = 0$ (number of evaluations performed)

$N = \max$ (number of HOO instances)

$S = \{(P_h,\rho_{max})\}$ (set of HOO instances)

while computational budget is available do

for $i = 1, \ldots, N$ do

$S \leftarrow \{(P_h,\rho_{max})\}$

(S) function evaluation with HOO(s)

Update the average reward $\bar{r}[i]$ of HOO(s)

for $n = 0$ do

$N = 2N$

end while

for $s \subseteq \{2\}$ do

Perform a function evaluation with HOO(s)

Update the average reward $\bar{r}[i]$ of HOO(s)

end for

Output: The deepest point evaluated by HOO(s)

**How it works?**

- Poo makes the use of HOO as a subroutine, an algorithm that requires the knowledge of the function smoothness.

- Poo automatically launchess several HOO instances in parallel for different smoothness $(\nu, \rho)$.

- At the end, Poo selects the instance $x^*$ which performed the best and returns the deepest point selected by this instance.

- Why it works?

  - From the analysis: few HOO instances are needed – $O(\ln n)$.
  - From the experiments: most of the evaluations are the same! Saving time by sharing information over HOO instances.

**Measure of Complexity**

Definition of the near-optimality dimension $d$

Definition 1. For any partitionning $P$, real $\nu > 0$ and $\rho \in (0, 1)$ verifying Assumption 1:

$$d((\nu, \rho)) = \frac{\ln \left(\sum_{x \in P} f(x) \right)}{\ln \left(\frac{1}{1 - \rho/2}\right)}$$

where $N_0(c)$ is the number of near-optimal cells $P_h$ of depth $h$ i.e. cells such that

$$\sup_{x \in P_h} f(x) = f(x^*) - \epsilon$$

- It measures how much information $P$ gives us about $f$. The hierarchical partitionning $P$ is the only prior information available.

- It is the size of the near-optimal set. This set is the cells that any algorithm would have to sample in order to discover the optimum.

- Examples of $d = 0$ functions.
  - Any function with same order upper and lower envelopes near its maximum for the standard partitioning.
  - $A > 0$ function for the standard partitioning.

Functions that behave differently in different dimensions have also $d > 0$.

**Comparing to previous assumptions**

**Previous work**

Assumption 1: There exists $\nu, \rho \in (0, 1)$ such that

$$\forall h \geq 0, \forall x \in P_h, f(x) \geq f(x^*) - \nu\rho^h.$$