General parallel optimization
WITHOUT a metric

Xuedong Shang, Emilie Kaufmann, Michal Valko

ALT - 23 March 2019
Black box optimization

Any gradient information? No! Also called zero-order optimization.
Black box optimization

Any gradient information? No! Also called zero-order optimization.
Black box optimization

Any gradient information? No! Also called zero-order optimization.
Black box optimization

Any gradient information? No! Also called zero-order optimization.
Black box optimization

Any gradient information? No! Also called zero-order optimization.
Black box optimization

How to choose the next query?
Black box optimization
Black box optimization

How to choose the next query?
Black box optimization

How to choose the next query?
Black box optimization

Assumption: Known Lipschitz constant? No!
Black box optimization

Assumption: Bayesian Gaussian priors? No!
Black box optimization

We want minimal assumptions! What is the smoothness of $f$?
Goal: Maximize $f : \mathcal{X} \rightarrow \mathbb{R}$ given a budget of $n$ evaluations.

Challenge: $f$ has an unknown smoothness.

Protocol: At round $t$, select $x_t$, observe $y_t$ such that

$$\mathbb{E}[y_t|x_t] = f(x_t) \quad |y_t - x_t| \leq 1$$

After $n$ rounds, return $x(n)$.

Loss: $r_n \triangleq \sup_{x \in \mathcal{X}} f(x) - f(x(n))$ (simple regret)
Minimal assumptions

- We want minimal assumptions.

- The smoothness of $f$ is defined with respect to a partitioning $\mathcal{P}$ of the search space $\mathcal{X}$.  **No metric!**  (Grill et al., 2015)
Minimal assumptions. Step 1. Partitioning

- For any depth $h$, $\mathcal{X}$ is partitioned in $K^h$ cells $(P_{h,i})_{0 \leq K^h - 1}$.
- $K$-ary tree $T$ where depth $h = 0$ is the whole $\mathcal{X}$.

An example of partitioning in one dimension with $K = 3$. 
Use the partitioning to explore \((\text{uniformly}) \, f\)
Optimizing becomes a tree search on the partition $\mathcal{P}$.

How to explore the tree smartly? (Track $x^*$ as deep as possible)
Tree search

Optimizing becomes a tree search on the partition \( \mathcal{P} \).

Opening a cell means evaluating all its \( K \) children cells.
The assumption and the smoothness

Assumption (Grill et al., 2015)

For some global optimum $x^*$, there exists $\nu > 0$ and $\rho \in (0, 1)$ such that $\forall h \in \mathbb{N}, \forall x \in \mathcal{P}_{h,i^*_h}$,

$$f(x) \geq f(x^*) - \nu \rho^h.$$

- The smoothness is local, around a $x^*$.
- This guarantees that the algorithm will not under-estimate by more than $\nu \rho^h$ the value of optimal cell $\mathcal{P}_{h,i^*_h}$ if it observes $f(x)$ with $x \in \mathcal{P}_{h,i^*_h}$.
- Now for the opposite question: How much non-optimal cells have values $\nu \rho^h$-close to optimal and therefore indiscernible from it? Let us **count** them!
The smoothness and the near-optimal dimension

**Definition**

For any $\nu > 0$, $C > 1$, and $\rho \in (0, 1)$, the *near-optimality dimension* $d(\nu, C, \rho)$ of $f$ with respect to the partitioning $\mathcal{P}$, is

$$d(\nu, C, \rho) \triangleq \inf \left\{ d' \in \mathbb{R}^+ : \forall h \geq 0, \mathcal{N}_h(3\nu \rho^h) \leq C \rho^{-d'h} \right\}.$$
The smoothness and the near-optimal dimension

**Definition**

For any $\nu > 0$, $C > 1$, and $\rho \in (0, 1)$, the *near-optimality dimension* $d(\nu, C, \rho)$ of $f$ with respect to the partitioning $\mathcal{P}$, is

$$
\boxed{
    d(\nu, C, \rho) \triangleq \inf \left\{ d' \in \mathbb{R}^+ : \forall h \geq 0, \mathcal{N}_h(3\nu \rho^h) \leq C \rho^{-d'h} \right\}.
}
$$

- $\mathcal{N}_h(\varepsilon)$ is the number of cells $\mathcal{P}_{h,i}$ of depth $h$ such that
  $\sup_{x \in \mathcal{P}_{h,i}} f(x) \geq f(x^*) - \varepsilon$.
- $\mathcal{N}_h(3\nu \rho^h)$ explodes exponentially w.r.t $d$. 


Previous work

Previous algorithms that depend on a metric:

<table>
<thead>
<tr>
<th>smoothness</th>
<th>global</th>
<th>local</th>
</tr>
</thead>
<tbody>
<tr>
<td>((v, \rho)) known</td>
<td>Zooming, HOO</td>
<td>DOO, HCT</td>
</tr>
<tr>
<td>((v, \rho)) unknown</td>
<td>TaxonomyZoom</td>
<td>StoSOO, SOO, ATB</td>
</tr>
</tbody>
</table>

We tackle **unknown** smoothness \((v, \rho)\) **without a metric**:

- **P00** (Grill et al., 2015) \(\mapsto\) requires a base algorithm that has upper-bounded **cumulative regret**
- **GPO** (our algorithm) \(\mapsto\) requires a base algorithm that has upper-bounded **simple regret**
The GPO algorithm

How it works?

→ We run several instances of the base algorithm over \( n/2 \).
The GPO algorithm

**Parameters:** base algorithm $\mathcal{A}$, $n$, $\mathcal{P} = \{\mathcal{P}_{h,i}\}$, $\rho_{\text{max}}$, $\nu_{\text{max}}$

**Initialization:** $D_{\text{max}} \leftarrow \ln K / \ln(1/\rho_{\text{max}})$

Compute $N = \lceil (1/2)D_{\text{max}} \ln((n/2)/\ln(n/2)) \rceil$

For $i = 1, \ldots, N$

1. $s \leftarrow (\nu_{\text{max}}, \rho_{\text{max}}^{2N/(2i+1)})$
2. Run $\mathcal{A}(s)$ for $\lceil n/(2N) \rceil$ time steps $\rightarrow \tilde{x}_s$

**Output**
The GPO algorithm

How it works?

⇝ We run several instances of the base algorithm over $n/2$.

⇝ We use another $n/2$ to do a cross-validation.
The GPO algorithm

**Parameters:** base algorithm $\mathcal{A}$, $n$, $\mathcal{P} = \{\mathcal{P}_{h,i}\}$, $\rho_{\text{max}}$, $\nu_{\text{max}}$

**Initialization:** $D_{\text{max}} \leftarrow \ln K / \ln(1/\rho_{\text{max}})$

Compute $N = \lceil (1/2)D_{\text{max}} \ln((n/2)/\ln(n/2)) \rceil$

**For** $i = 1, \ldots, N$

- $s \leftarrow (\nu_{\text{max}}, \rho_{\text{max}}^{2N/(2i+1)})$
- Run $\mathcal{A}(s)$ for $\lfloor n/(2N) \rfloor$ time steps $\rightarrow \tilde{x}_s$
- Get $\lfloor n/(2N) \rfloor$ evaluations of $f(\tilde{x}_s) \rightarrow$ average $V[s]$

$s^* \leftarrow \arg \max_s V[s]$

**Output** $x(n) \leftarrow \tilde{x}_{s^*}$
The GPO algorithm

Theorem

If for all \((\nu, \rho)\) the \(A(\nu, \rho)\) algorithm has its simple regret bounded as

\[
\mathbb{E}\left[r_n^A(\nu, \rho)\right] \leq \alpha C \left(\frac{\log n}{n}\right)^{1/(d+2)},
\]

for any function \(f\) satisfying our minimal assumptions with parameters \((\nu, \rho)\), then there exists a constant \(\beta\) that is independent of \(\nu_{\text{max}}\) and \(\rho_{\text{max}}\) such that

\[
\mathbb{E}\left[r_n^{\text{GPO}(A)}\right] \leq \beta D_{\text{max}} \left(\frac{\nu_{\text{max}}}{\nu^*}\right)^{D_{\text{max}}} \left(\frac{\log^2 n}{n}\right)^{1/(d+2)},
\]

for any function \(f\) satisfying our minimal assumptions with parameters \(\nu^* \leq \nu_{\text{max}}\) and \(\rho^* \leq \rho_{\text{max}}\).
The GPO algorithm

How it works?

⇝ The question now is whether there exists a base algorithm that has simple regret guarantee (1) under our minimal assumptions.
How it works?

⇝ The question now is whether there exists a base algorithm that has simple regret guarantee (1) under our minimal assumptions.

⇝ It is not clear whether HOO satisfies our needs. Worse, it is not even clear that HOO has any regret bound under our minimal assumptions, contrary to what is claimed by Grill et.al. (2015).
The GPO algorithm

How it works?

⇝ The question now is whether there exists a base algorithm that has simple regret guarantee (1) under our minimal assumptions.

⇝ It is not clear whether HOO satisfies our needs. Worse, it is not even clear that HOO has any regret bound under our minimal assumptions, contrary to what is claimed by Grill et.al. (2015).

⇝ HCT does! (Azar et.al., 2014): requires a refined analysis.
The simple regret of HCT after $n$ function evaluations under our minimal assumptions satisfies

$$E[r_n^{HCT(\nu, \rho)}] \leq \alpha C \left( \frac{\log n}{n} \right)^{1/(d+2)}.$$
Takeaway messages:

- A general meta-algorithm that adapts to unknown local smoothness that only requires the base algorithm to have some simple regret guarantee.
- Refined HCT analysis showing that it is a valid candidate.

Thank you!