Derivative-Free Optimization via Classification

Yang Yu, Hong Qian and Yi-Qi Hu

Nanjing University

May 25, 2017

Presented by Yifan Ge

Section 1

Background

Derivative-Free Optimization

- $\text{argmin}_{x \in X} f(x)$
- ~~linearity, convexity, **differentiability**~~
- genetic algorithms, randomized local search, estimation of distribution algorithms, cross-entropy methods, Bayesian optimization methods, optimistic optimization methods, etc.
- usually model-based

Classification-based Optimization

Algorithm 1 classification-based optimization

Input:
f: Objective function to be minimized;
$\mathcal{C}$: A binary classification algorithm;
$\lambda \in [0,1]$: Balancing parameter;
$\alpha_1 > \ldots > \alpha_T$: Threshold for labeling;
$T \in \mathbb{N}^+$: Number of iterations;
$m \in \mathbb{N}^+$: Sample size in each iteration;
`Sampling`: Sampling sub-procedure.

Procedure:

```
Collect S_0 = {x_1, ..., x_m} by i.i.d. sampling from U_X
Let x~ = argmin_{x ∈ S_0} f(x)
for t = 1 to T do
    Construct B_t = {(x_1, y_1), ..., (x_m, y_m)},
    where x_i ∈ S_{t-1} and y_i = sign[α_t − f(x_i)]
    Let S_t = ∅
    for i = 1 to m do
        h_t = C(B_t), where h_t ∈ H
        x_i = Sampling(h_t, λ), and let S_t = S_t ∪ {x_i}
    end for
    x~ = argmin_{x ∈ S_t ∪ {x~}} f(x)
end for
return x~ and f(x~)
```

- Let sign$[v]$ be the sign function returning 1 if $v \geq 0$ and 1 otherwise.
- We specify the `Sampling`(h, λ) as that, it samples with probability λ from $\mathcal{U}_{D_h}$ (the uniform distribution over the positive region classified by h), and with the remaining probability from $\mathcal{U}_X$ (the uniform distribution over X).

Section 2

Theoretical Study

(ϵ, δ)-Query Complexity

Given $f \in \mathcal{F}$, an algorithm A, $0 < \delta < 1$ and $\epsilon > 0$, the (ϵ, δ)-query complexity is the number of calls to f such that, with probability at least $1 - \delta$, A finds at least one solution $\tilde{x} \in X \subseteq \mathbb{R}^n$ satisfying

$$f(\tilde{x}) - f(x^*) \leq \epsilon\,,$$

where $f(x^*) = \min_{x \in X} f(x)$.

Section 3

RACOS

Randomized Coordinate Shrinking

Algorithm 2 The *randomized coordinate shrinking* classification algorithm for $X = \{0,1\}^n$ or $[0,1]^n$

Input:

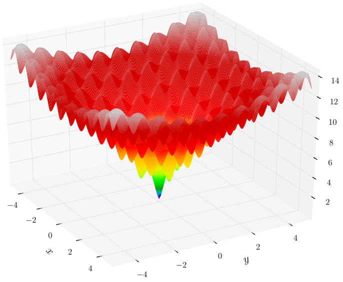
t: Current iteration number;
B_t: Solution set in iteration t;
X: Solution space ($\{0,1\}^n$ or $[0,1]^n$);
I: Index set of coordinates;
$M \in \mathbb{N}^+$: Maximum number of uncertain coordinates.

Procedure:

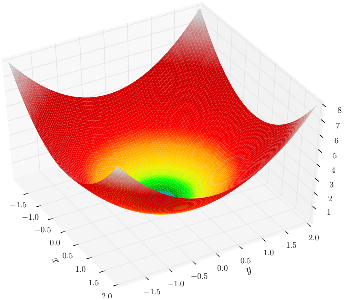
B_t^+ = the positive solutions in B_t
$B_t^- = B_t - B_t^+$
Randomly select $x_+ = (x_+^{(1)}, \ldots, x_+^{(n)})$ from B_t^+
Let $D_{h_t} = X$, $I = \{1, \ldots, n\}$
while $\exists x \in B_t^-$ s.t. $h_t(x) = +1$ **do**
  if $X = \{0,1\}^n$ **then**
    k = randomly selected index from the index set I
    $D_{h_t} = D_{h_t} - \{x \in X \mid x^{(k)} \neq x_+^{(k)}\}$, $I = I - \{k\}$
  end if
  if $X = [0,1]^n$ **then**
    k = randomly selected index from the index set I
    x^- = randomly selected solution from B_t^-
    if $x_+^{(k)} \geq x_-^{(k)}$ **then**
      r = uniformly sampled value in $(x_-^{(k)}, x_+^{(k)})$
      $D_{h_t} = D_{h_t} - \{x \in X \mid x^{(k)} < r\}$
    else
      r = uniformly sampled value in $(x_+^{(k)}, x_-^{(k)})$
      $D_{h_t} = D_{h_t} - \{x \in X \mid x^{(k)} > r\}$
    end if
  end if
end while
while $\#I > M$ **do**
  k = randomly selected index from the index set I
  $D_{h_t} = D_{h_t} - \{x \in X \mid x^{(k)} \neq x_+^{(k)}\}$, $I = I - \{k\}$
end while
return h_t

- For a subset $D \subseteq X$, let $\#D = \int_{x \in X} \mathbb{I}[x \in D] \mathrm{d}x$ (or $\#D = \sum_{x \in X} \mathbb{I}[x \in D]$ for finite discrete domains), where $\mathbb{I}[\cdot]$ is the indicator function.
- $D_h = \{x \in X \mid h(x) = +1\}$.

Ackley Function & Sphere Function



(a) Ackley function for $n = 2$



(b) Sphere function for $n = 2$

1

[1]Wikipedia

Ackley Function & Sphere Function

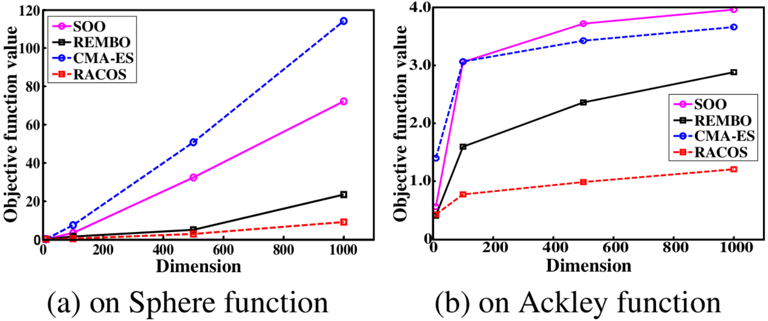


Figure: Comparing the scalability with $30n$ evaluations

Ackley Function & Sphere Function

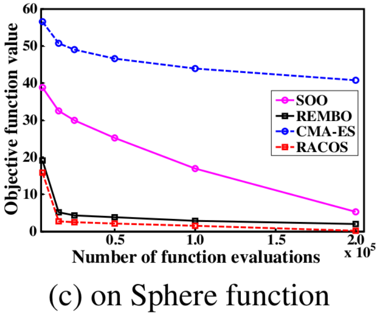


Figure: Comparing the convergence rate with $n = 500$

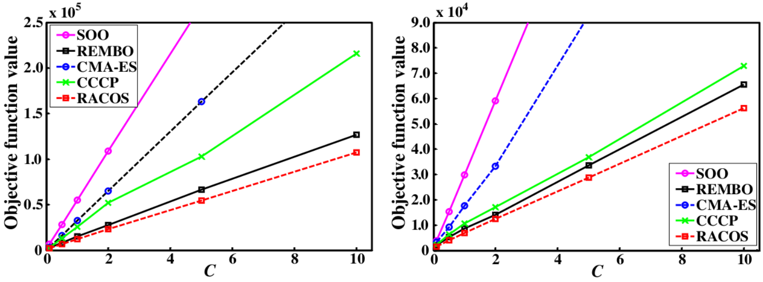
Classification with Ramp Loss

- Hinge Loss: $H_s(z) = \max 0, s - z$.
- Ramp Loss: $R_s(z) = H_1(z) - H_s(z)$, $s < 1$.
- Objective Function:

$$f(w, b) = \frac{1}{2}\|w\|_2^2 + C\sum_{l}^{L} R_s(y_l(w^T v_l + b)).$$

- NON-CONVEX!

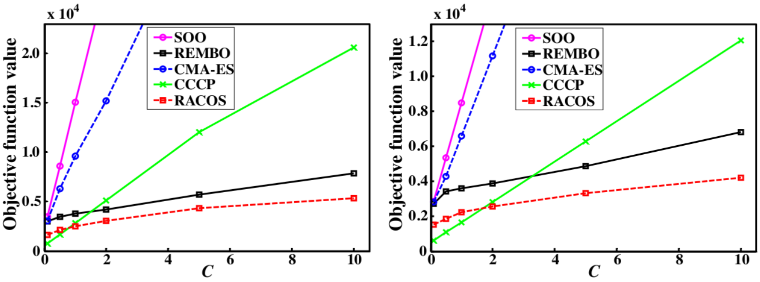
Classification with Ramp Loss



(a) on *Adult*, $s = -1$ (b) on *Adult*, $s = 0$

Figure: Comparing the achieved objective function values with $40n$ evaluations against the parameter C of the classification with Ramp loss.

Classification with Ramp Loss



(c) on *USPS+N*, $s = -1$ (d) on *USPS+N*, $s = 0$

Figure: Comparing the achieved objective function values with $40n$ evaluations against the parameter C of the classification with Ramp loss.