Evolutionary Gradient Descent for Non-convex Optimization

Ke Xue[1], Chao Qian[1*], Ling Xu[2], Xudong Fei[2].
{xuek, qianc}@lamda.nju.edu.cn
[1]LAMDA Group, Nanjing University, China
[2]Huawei Technologies, China

IJCAI 2021 MONTREAL

Background and Motivation

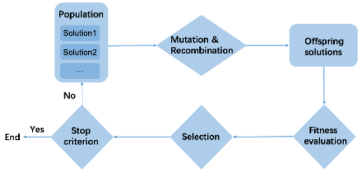
Non-convex optimization

- Popular in many real-world tasks
- Hard to solve
  - First order stationary point is global minima in convex optimization.
  - However, it maybe saddle point in non-convex optimization.

How to efficiently escape saddle points and find second order stationary point is the key issue in non-convex optimization.

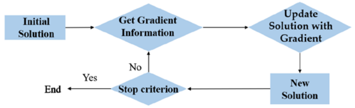
Evolutionary Algorithms

- Global convergence
- Low efficiency, especially in high dimension



Gradient Descent

- Perform well in high dimension and large scale tasks
- Converge to local optima generally



EAs and GD each has their advantages and disadvantages.

Can we get better algorithm for non-convex optimization by combining the merits of EAs and GD?

EGD Algorithm

Algorithm 2 EGD algorithm

Parameter: learning rate η, population size N, mutation strength $\{r^{(p)}\}_{p=1}^N$, time interval L for mutation, tolerance ϵ, ϵ', number T of iterations

Process:

Initialize the population $\{\boldsymbol{x}_0^{(p)}\}_{p=1}^N$, set $i = i_{\text{mutate}} = 0$, $update^{(p)} = 1$ for $p \in [N]$;
while $i \le T$ **do**
  for $p = 1 : N$ **do**
    if $update^{(p)} = 1$ **then**
      if $\|\nabla f(\boldsymbol{x}_i^{(p)})\| \le \epsilon$ and $i - i_{\text{mutate}} > L$ **then**
        $update^{(p)} = 0$
      else
        $\boldsymbol{x}_{i+1}^{(p)} \leftarrow \boldsymbol{x}_i^{(p)} - \eta \nabla f(\boldsymbol{x}_i^{(p)})$
      end if
    end if
  end for
  if $\forall p \in [N] : update^{(p)} = 0$ **then**
    Apply Algorithm 3 for mutation and selection;
    Set $update^{(p)} = 1$ for $p \in [N]$;
    $i \leftarrow i + L$
  end if
  $i \leftarrow i + 1$
end while

Gradient descent update or wait for mutation

Algorithm 3 Mutation and Selection

for $p = 1 : N$ **do**
  $\boldsymbol{x}_{i+1}^{(p)} \leftarrow \boldsymbol{x}_i^{(p)} + \boldsymbol{\xi}_i^{(p)}$, $\boldsymbol{\xi}_i^{(p)} \sim \text{Uniform}(B(\mathbf{0}, r^{(p)}))$;
  $i_{\text{mutate}} \leftarrow i$;
  for $j = (i+1) : (i+L)$ **do**
    $\boldsymbol{x}_{j+1}^{(p)} \leftarrow \boldsymbol{x}_j^{(p)} - \eta \nabla f(\boldsymbol{x}_j^{(p)})$
  end for
  if $f(\boldsymbol{x}_{i+1+L}^{(p)}) + \epsilon' < f(\boldsymbol{x}_i^{(p)})$ **then**
    $escape^{(p)} = 1$
  else
    $escape^{(p)} = 0$, $\boldsymbol{x}_{i+1+L}^{(p)} \leftarrow \boldsymbol{x}_i^{(p)}$
  end if
end for
$f_{\text{mean}} = \frac{1}{N}\sum_{p=1}^N f(\boldsymbol{x}_{i+1+L}^{(p)})$;
$\boldsymbol{x}_{\text{best}} = \arg\min_{\boldsymbol{x} \in \{\boldsymbol{x}_{i+1+L}^{(p)}\}_{p=1}^N} f(\boldsymbol{x})$;
for $p = 1 : N$ **do**
  if $escape^{(p)} = 0$ and $f(\boldsymbol{x}_{i+1+L}^{(p)}) \ge f_{\text{mean}}$ **then**
    $\boldsymbol{x}_{i+1+L}^{(p)} \leftarrow \boldsymbol{x}_{\text{best}}$
  end if
end for

Mutation and update for *L* iterations

Selection

Theoretical Analysis

Theorem 2. *Let f satisfies Assumption 1. Let the parameters of EGD satisfy that $\eta = \frac{1}{\ell}$, $L = \frac{\ell}{\sqrt{\rho\epsilon}} \cdot \iota$ and $\epsilon' = \frac{1}{100\iota^3}\sqrt{\frac{\epsilon^3}{\rho}}$, where $\iota = c\log(\frac{d\ell(f(\boldsymbol{x}_0)-f^*)}{\rho\epsilon\delta})$, and c is an absolute constant. Then for any $\epsilon, \delta > 0$, after running*

$$\tilde{O}(\ell(f(\boldsymbol{x}_0) - f^*)/\epsilon^2)$$

iterations, EGD will find an ϵ-second-order stationary point with probability at least $1 - \delta_{egd}$, where

$$\delta_{egd} = \frac{T^*}{4L} \cdot \prod_{p=1}^{N} \frac{\ell\sqrt{d\epsilon}}{r^{(p)}\sqrt{\rho}} \frac{1}{\sqrt{\pi}2^{\iota}\iota},$$

and $T^* = 8\max\{50\ell(f(\boldsymbol{x}_0) - f^*) \cdot \iota^4, \ell(f(\boldsymbol{x}_0) - f^*)\}/\epsilon^2$.

Theorem 2 give the iterations and probability of EGD to find ϵ-second-order stationary point.

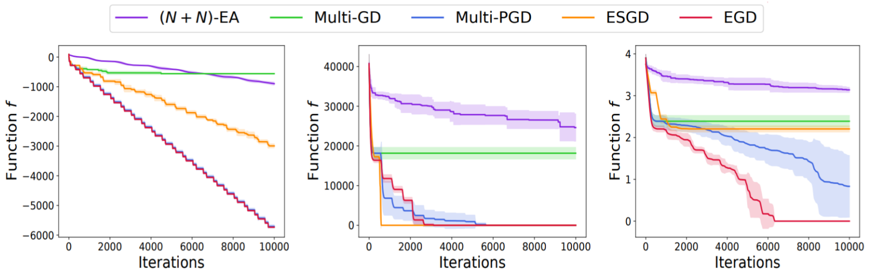
Remark 1. *EGD will have more advantage over Multi-PGD,*

(1) *when the problem dimension d, the Lipschitz parameters ℓ and ρ are larger, implying that the problem is more challenging;*

(2) *when the population size N is larger.*

Remark2 shows that when EGD is much better than baseline methods.

Experiment



(a) *Octopus* (b) *Schwefel* (c) *Ackley*

Dimension d	200	400	600	800	1000
$f(\boldsymbol{x}) < 2$	1.79	2.05	2.71	2.78	2.88
$f(\boldsymbol{x}) < 1$	1.70	2.04	2.24	2.30	2.34
$f(\boldsymbol{x}) < 0.1$	1.51	1.88	2.10	2.20	2.24

the ratio of the number of iterations of Multi-PGD and EGD until finding a solution with the value smaller than a threshold