Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2021-0144

Vol. **31**, No. 4, pp. 1296-1316
April 2022

A Consensus-Based Global Optimization Method with Adaptive Momentum Estimation

Jingrun Chen[1,2], Shi Jin[3] and Liyao Lyu[4,*]

[1] *School of Mathematical Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China.*
[2] *Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215123, China.*
[3] *School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai, 200240, China.*
[4] *Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, 48824, USA.*

Received 5 July 2021; Accepted (in revised version) 15 February 2022

Abstract. Objective functions in large-scale machine-learning and artificial intelligence applications often live in high dimensions with strong non-convexity and massive local minima. Gradient-based methods, such as the stochastic gradient method and Adam [15], and gradient-free methods, such as the consensus-based optimization (CBO) method, can be employed to find minima. In this work, based on the CBO method and Adam, we propose a consensus-based global optimization method with adaptive momentum estimation (Adam-CBO). Advantages of the Adam-CBO method include:

- It is capable of finding global minima of non-convex objective functions with high success rates and low costs. This is verified by finding the global minimizer of the 1000 dimensional Rastrigin function with 100% success rate at a cost only growing linearly with respect to the dimensionality.
- It can handle non-differentiable activation functions and thus approximate low-regularity functions with better accuracy. This is confirmed by solving a machine learning task for partial differential equations with low-regularity solutions where the Adam-CBO method provides better results than Adam.
- It is robust in the sense that its convergence is insensitive to the learning rate by a linear stability analysis. This is confirmed by finding the minimizer of a quadratic function.

AMS subject classifications: 37N40, 90C26
Key words: Consensus-based optimization, global optimization, machine learning, curse of dimensionality.

*Corresponding author. *Email addresses:* jingrunchen@ustc.edu.cn (J. Chen), shijin-m@sjtu.edu.cn (S. Jin), lyuliyao@msu.edu (L. Lyu)

1 Introduction

The goal of this work is developing consensus-based global optimization methods to solve high dimensional unconstrained optimization problems

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} f(\theta),$$

where the target function (loss function) $f(\theta)$ defined in $\mathbb{R}^d$ achieves a unique global minimizer.

A high-dimensional nonlinear, non-convex optimization is an essential part of machine learning problems, with the target function defined in general as

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \|\mathcal{N}_\theta(\hat{x}_i) - \hat{y}_i\|,$$

where θ is the parameter vector and $\mathcal{N}_\theta$ represents a neural network representation. $(\hat{x}_i,\hat{y}_i)_{i=1}^n$ is a set of labeled data, and $\|\cdot\|$ is the L^2 distance between a predicted data point and the corresponding labeled data point.

The gradient descent method, the most frequently used method in optimization, often updates the parameters by the iteration scheme

$$\theta^{t+1} = \theta^t - \alpha\nabla f(\theta^t),$$

where θ^0 is initialized by a normal distribution with the mean and the variance specified in [9, 13] and α is the learning rate. However, for a big labeled data set, i.e., n is tremendously big, computing f in each iteration is time consuming, and the iterations often get stuck at local minima. The stochastic gradient descent (SGD) method [2, 3] approximates f by

$$\hat{f}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \|\mathcal{N}_\theta(\hat{x}_i) - \hat{y}_i\|$$

on a randomly selected subset of the labeled data set, by choosing m points randomly from the labeled data set with $m \ll n$. Note that the subset needs to be updated at each iteration.

The SGD method with momentum term [20] damps oscillations in the SGD method by introducing exponentially weighted moving average as the momentum

$$\begin{aligned}\theta^{t+1} &= \theta^t - m^t,\\ m^t &= \gamma m^{t-1} + \alpha\nabla_\theta \hat{f}(\theta^t).\end{aligned}$$

The initialization of θ^0 is the same as that in SGD and the momentum m^0 is initialized to be zero. In the component-wise sense, the momentum term increases for dimensions whose gradients point toward the same direction and decreases for dimensions whose