On the global minimum convergence of non-convex deterministic functions via Stochastic Approximation

Charlie Chen
Courant Institutte of Mathematical Science
New York University
New York, NY, 10003
charlie.chen@nyu.edu

Guanming Zhang
Center for Soft Matter Research
New York University
New York, NY, 10003
gz2241@nyu.edu

Stefano Martiniani
Center for Soft Matter Research (CSMR), Courant Institutte of Mathematical Science
New York University
New York, NY, 10003
stefano.martiniani@nyu.edu

Abstract

Guarantee to converging to global minimum is hard to obtain, except for convex cases. Recently, Caravelli *et al*. [2] proposed an optimization algorithm that converges to the global minimum, but the reason of convergence is not clear. In this paper, we argue that the algorithm can be approximated by optimizing a distribution on the objective function. We further propose a new algorithm named SA-PEDS and show that it performs better than the original algorithm on the Ackley function.

1 Introduction

The common form of unconstrained optimization is

$$\min_{X\in\mathbb{R}^m} f(X), \tag{1}$$

where m is the dimension of the original problem. If f is a convex function, a lot of algorithms like Newton's Method have guaranteed convergence bound[4]. However, for non-convex objective functions, convergence to global minimum cannot be guaranteed in general[5]. Luckily, ensuring global convergence for all functions is not what we want, as most optimization theories are developed when the objective functions belong to certain hypothesis set. It is possible to provide a global convergence theory under this restriction.

In the exploration of memristors, a circuit of memristors optimize their internal states to achieve global minimum, even when it is initialized close to some local minimum [1]. In a following work, they proposed an optimization algorithm, Projective Embedding of Dynamical Systems (PEDS), which embeds the original m-dimensional problem to mN-dimensional space while using some projection matrix to limit its degree of freedom[2]. This algorithms showcases capability to find the global minimum, but the reason for convergence is not clear. To better understand this algorithm, we present a new way to view PEDS when some particular configuration is chosen. Instead finding one minimizer, the new approach utilizes distribution as the variable to learn in the training phase. The distribution converges to a point (delta function) at the end of the training. This provides new inspirations for optimization problems.

In this paper, after summarizing this PEDS algorithm, we will show that under some particular case, the algorithm can be seen as a Stochastic Approximation algorithm and we will show that the new

Applied Math Summer Undergraduate Research Experience (AM-SURE) 2023

algorithm performs better than the original problem for the Ackley test function, especially when the dimension of the problem is high.

2 Related Work

In this section, we are going to show that a particular case of PEDS can be seen as a particle algorithm with some special interaction force among them. First, we formulize the algorithm. Given the problem as stated in Eq. 1, extend the variable to $M \in \mathbb{R}^{N\times m}(\{m_{i,j}\}_{i=1,\dots,N;j=1,\dots,m})$. Denote the column vector as Y_j. Then, the update is described by

$$Y_j^{t+1} - Y_j^t = -\gamma(\Omega\, \Phi(\nabla F; Y_1^t, Y_2^t, \dots, Y_m^t) + \alpha(I-\Omega)Y_j^t), \tag{2}$$

where Φ is matrix map, Ω is projection matrix ($\Omega^2 = \Omega$), and γ and α are some hyperparameters.

For a particular choice of Φ, we define it as

$$\Phi(\nabla F; Y_1, Y_2, \dots Y_m)_i = \nabla F\left((m_{i,1}, m_{i,2}, \dots m_{i,m})^T\right) = \nabla F(R_i), \tag{3}$$

where R_i is the ith row of the matrix M. For the projection matrix, we have $\Omega = \Omega_1 = \frac{1}{N}\begin{pmatrix} 1\cdots 1 \\ \vdots\ddots\vdots \\ 1\cdots 1 \end{pmatrix}$, called mean-field projector.

With these two choice and through some straightforward algebra, we see that the update can be written as

$$R_i^{t+1} - R_i^t = -\gamma\left(\frac{1}{N}\sum_{i=1}^N \nabla F(R_i^t) + \alpha(R_i^t - \overline{R}^t)\right), \tag{4}$$

where $\overline{R} = \frac{1}{N}\sum_{i=1}^N R_i$. Think of each R_i as a particle and at each update, they update according to the average of all gradients and another attraction force that pulls all particles to the center of mass.

3 Method

Inspired by the formulation in Eq. 4, we design so-called Stochastic Approximation Projective Embeddings of Dynamical Systems (SA-PEDS). To see this, we will take a closer look at Eq. 4. Instead of treating the independent variable as a fixed variable, but rather as samples from some distribution. Now we optimize over the distribution. The problem is then formulated as

$$\min_{\mathcal{D}\in\Delta(\mathbb{R}^m)} \mathbb{E}f(X) \tag{5a}$$

$$\text{s.t.} \quad X \sim \mathcal{D} \tag{5b}$$

where $\Delta(\mathbb{R}^m)$ is all distribution over $\mathbb{R}^m$.

When $\mathcal{D}$ is limited to the family of delta distribution, this recovers the original problem as stated in Eq. 1. It is also not hard to see, if the objective f has a unique global minimizer, the solution for Eq. 5 is exactly a delta distribution. In other words, we have

$$\min f(x) = \min_{\mathcal{D}\subset\{\delta(x):x\in\mathbb{R}\}} \mathbb{E}f(X) = \min_{\mathcal{D}\in\Delta(\mathbb{R}^m)} \mathbb{E}f(X). \tag{6}$$

Instead of limiting $\mathcal{D}$ to be delta distribution, we want to approximate it. In particular, recall that $\mathcal{N}(\theta, 0) = \delta(\theta)$. So it is equivalent to formulate the problem as

$$\lim_{\sigma\to 0}\min_{\theta} \mathbb{E}f(X), \tag{7a}$$

$$\text{s.t.} \quad X \sim \mathcal{N}(\theta, \sigma^2) \tag{7b}$$

which can be solved by a Stochastic Approximation algorithm.

We will see later why the Gaussian distribution approximation is helpful. Compare this with the formulation in Eq. 4. Observe that the first part is just the empirical evaluation of the gradient $\mathbb{E}\nabla f(X)$, if R_i is sampled from the target distribution. The algorithm is described in Algorithm 1.

Algorithm 1 Stochastic Approximation Projective Embeddings of Dynamical System (SA-PEDS)

Require: γ: Stepsize
Require: N: Number of samples
Require: $f(X)$: Objective function
Require: Optim: Optimizer
Require: α: Linear decreasing rate for standard deviation
Require: θ_0: Initial parameter vector
Require: σ_0: Initial variance of the Gaussian distribution
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
  $t \leftarrow t + 1$
  $R_1, \ldots, R_N \sim \mathcal{N}(\theta_{t-1}, \sigma_{t-1})$ (Draw samples from the normal distribution)
  $g_i \leftarrow \nabla_\theta f(R_i)$ (Compute the gradient for each sample)
  $g_t \leftarrow \frac{1}{N}\sum_{i=1}^{N} g_i$ (Calculate the average)
  $\theta_t \leftarrow \mathrm{optim}(\theta_{t-1}, \gamma, g_t)$ (Update the parameter)
  $\sigma_t \leftarrow \max(\sigma_{t-1} - \alpha, 0)$ (Update the std)
end while
return θ_t (Mean of the resulting distribution)

Here are some intuitions for the method. If R_i is sampled from $\mathcal{N}(\theta, \sigma)$. We have

$$\mathbb{E}\nabla F(R) = \int \nabla F(R)\mathcal{N}(R; \theta, \sigma^2 I)dR = \int \nabla F(R)\rho(\theta - R)dX = \nabla F * \rho(\theta), \tag{8}$$

where $\rho(X) \approx e^{-\|X\|^2}$ (up to some constants). We know that the convolution is at least as smooth as one of the convoluted functions, in this case, the Gaussian density function. The smoother the Gaussian density function is, i.e. the larger σ is, the smoother the expectation of gradient will be. This method is also used in the context of non-smooth Stochastic Gradient Descent, called Randomized Smoothing [3].

4 Experiments

In this section, we present the experiment results. The test function is Ackley function, as defined by

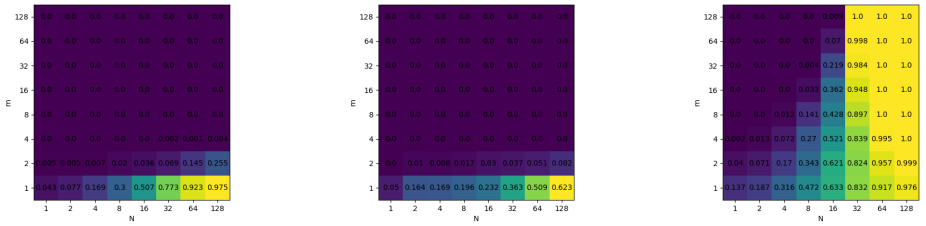
$$F(X) = -a\exp -b\sqrt{\frac{1}{m}\sum_{i=1}^{m} X_i^2} - \exp\frac{1}{m}\sum_{i=1}^{m}\cos(cX_i) + a + \exp 1, \tag{9}$$

where we use $a = 20, b = 0.2, c = 2\pi$. There are three methods: restart (take different initial values and optimize to some local minimum), PEDS (the algorithm described in Eq. 4, and SA-PEDS (the algorithm in Algorithm. 1. We are interested in two variables: success rate (the probability of converging to the global minimum) and convergence time (how long does the convergence take). We denote m for the dimension of Ackley function and N as the number of particles. The codes for the experiments can be accessed in `https://github.com/charliezchen/SA-PEDS`. The result of the experiment is shown in Figure. 1 and 2.

5 Conclusion

In conclusion, inspired by PEDS, we proposed SA-PEDS, which achieves successful convergence behavior on the Ackley function. Note that in the derivation of SA-PEDS, we pick a particular case in PEDS which is not true in general. It is still interesting to investigate why PEDS is an effective method in finding the global minimum.

In the analysis part of the algorithm, we see that SA-PEDS filters out the high frequency content of the signal and keep the lower frequency signal. This works if the high frequency part is noise and the lower frequency part is useful information. If it is the other way around, it is not clear why this method should work. This is part of the future work: testing this algorithm on more test functions. Moreover, it is also interesting to investigate on the optimal schedule of α, which affects the convergence results in the numerical experiments.



(a) (b) (c)

Figure 1. Success rate for different methods: (a) Restart (b) PEDS (c) SA-PEDS. Both algorithms perform well for low value of m, but for larger N, Restart and PEDS don't work anymore while SA-PEDS still work.

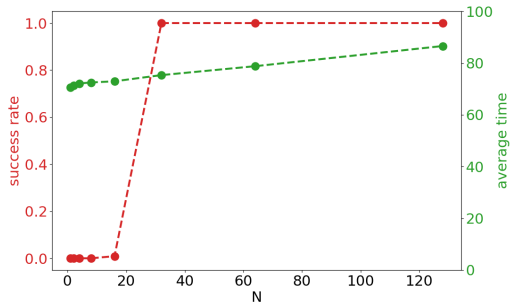


Figure 2. Convergence time for SA-PEDS. The time scales linearly with the increase of N, but the success rate has a jump, so the cost is acceptable. Furthermore, the algorithm can be accelerated using ideas of importance sampling and sliding windows.

References

[1] Francesco Caravelli, Forrest C Sheldon, and Fabio L Traversa. Global minimization via classical tunneling assisted by collective force field formation. *Science Advances*, 7(52):eabh1542, 2021. 1

[2] Francesco Caravelli, Fabio L Traversa, Michele Bonnin, and Fabrizio Bonani. Projective embedding of dynamical systems: Uniform mean field equations. *Physica D: Nonlinear Phenomena*, 450:133747, 2023. 1

[3] John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012. 3

[4] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999. 1

[5] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997. 1