Acceleration

Jinxiong Zhang

March 2019

1 Introduction

It is popular to use Nesterov's accelerated gradient method to minimize the cost function $f(x)$ in machine learning specially deep learning, where it is called momentum method. The source of its acceleration is studied from many perspectives such as the blogs at `https://blogs.princeton.edu/imabandit/2015/06/30/revisiting-nesterovs-acceleration/`. If Anderson acceleration for fixed point iteration is applied to the problem

$$x = x - \alpha \nabla f(x)$$

where α is a scalar function, it is still an open problem whether it is equivalent to Nesterov's accelerated gradient method as far as known.

It is worthy of exploring it.

2 Nesterov's accelerated gradient method

The general form of Nesterov's accelerated gradient method to minimize the cost function $f(x)$ can be written in the following form:

$$x^{k+1} = x^k - \gamma \nabla f(x^k + \mu(x^k - x^{k-1})) + \mu(x^k - x^{k-1}).$$

The parameters γ and μ are difficult to tune when $f(x)$ is non-convex.

3 Anderson acceleration

We apply Anderson acceleration to the problem $x = x - \alpha \nabla f(x)$:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + (1-\alpha_1)\alpha[\nabla f(x^k) - \nabla f(x^{k-1})] + (1-\alpha_1)(x^{k-1} - x^k)$$

where $\alpha_1 = \arg\min_{\alpha_1} \|\alpha_1 \nabla f(x^k) + (1-\alpha_1)\nabla f(x^{k-1})\|_2^2 = -\frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|_2^2}{\langle \nabla f(x^k) - \nabla f(x^{k-1}), \nabla f(x^{k-1})\rangle}$.

There is a difference of previous iteration in Anderson acceleration as well as Nesterov's accelerated gradient method. It seems that

$$\gamma \nabla f(x^k + \mu(x^k - x^{k-1})) \approx (1-\alpha_1)\alpha[\nabla f(x^k) - \nabla f(x^{k-1})].$$

And what is the connection of these two methods in mathematics?