Quick Review

1. Mathematicians have proved that in order to achieve linear convergence, we must have a function with strong convexity and Lipschitz smoothness:

$$\underbrace{\frac{\mu}{2}\|x-y\|_2^2}_{\text{strong convexity}} \le f(x)-f(y)-\langle \nabla f(y), x-y\rangle \le \underbrace{\frac{L}{2}\|x-y\|_2^2}_{\text{smoothness}} \qquad (\mu, L>0)$$

linear convergence

if $\lim_{n\to\infty} x_n = L$ (limit is L)

then $\lim_{n\to\infty} \frac{|x_{n+1}-L|}{|x_n-L|} = \mu$

if $0<\mu<1$, linear

if $\mu=0$, super linear

if $\mu=1$, sub-linear

convexity

$\begin{cases} f''(x)\ge 0 & \text{convex} \\ f''(x)>0 & \text{strictly convex} \\ f''(x)\ge m>0 & \text{strongly convex} \end{cases}$

2. In general, we know:

$\begin{cases} \text{smoothness + strong convexity + first order method} \longrightarrow \text{linear convergence} \\ \text{smoothness + convexity + first order method} \longrightarrow \text{sub-linear or worse} \\ \text{strong convexity + first order method} \longrightarrow \text{sub-linear or worse} \end{cases}$

3. In this paper, we break this rule by relaxing the lower bound:

$$m\|x-x_{min}\|^b \le f(x)-f(x_{min}) \le M\|x-x_{min}\|^b \qquad (b>1,\ m\ge 0,\ M>0)$$

which means you don't have to be strongly convex (you just have to be convex). And of course, you also need to check some other restrictions in the paper based on different conditions! Note that strong convexity and smoothness is more restrictive and they only cover $b=2$ case. If $b\neq 2$, gradient descent does poorly in general.

4. In fact, the "trick" of this paper is to use the gradients of two functions (∇f and ∇k) instead of one (∇f) to achieve exponential convergence.

5. Motivated by polyak's heavy ball method, the system is actually the generalization of heavy ball method. (kinetic energy term to be quadratic = standard momentum)

Heavy ball method

$$\begin{cases} x_t' = p_t \\ p_t' = -\nabla f(x_t) - \gamma p_t \end{cases}$$

Heavy ball described in Hamiltanian System in first order derivative

identical

Momentum Gradient Method (MGM)

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \rho (x_k - x_{k-1})$$

η: step size ...

$\rho (x_k - x_{k-1})$: Polyak's momentum

ρ is a momentum parameter

if $\rho = 0$, MGM becomes Vanilla Gradient Method (VGM)

also known as Gradient descent.

Heavy ball described in its discrete approximation.