Exact rate of Nesterov Scheme

Vasileios Apidopoulos, Jean-François Aujol, Charles Dossal, Aude Rondepierre

INSA de Toulouse, Institut de Mathématiques de Toulouse

January 2019, MIA Conference

Minimize a differentiable function

Let F be a convex differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$ which gradient is $L - Lipschitz$, having at least one minimizer x^*.
We want to build an efficient sequence to estimate

$$\arg\min_{x \in \mathbb{R}^n} F(x) \tag{1}$$

Explicit Gradient Descent

Let F be a convex differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$ which gradient is $L-Lipschitz$, having at least one minimizer x^*.

- Gradient descent : for $h < \frac{2}{L}$,

$$x_{n+1} = x_n - h\nabla F(x_n) \tag{2}$$

The sequence $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer of F and

$$F(x_n) - F(x^*) \leqslant \frac{\|x_0 - x^*\|^2}{2hn} \tag{3}$$

Nesterov inertial scheme

- Nesterov Scheme for $h < \frac{1}{L}$, and $\alpha \geqslant 3$

$$x_{n+1} = x_n - h\nabla F\left(x_n + \frac{n}{n+\alpha}(x_n - x_{n-1})\right) \tag{4}$$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^2}\right) \tag{5}$$

- Nesterov (84) proposes $\alpha = 3$.

The questions

- More precise than $O\left(\frac{1}{n^2}\right)$ with more information on F ?
- Is Nesterov really an acceleration of Gradient descent ?

The answers

- Yes... with strong convexity, Su et al. (15) Attouch et al. (17)
- We give a more accurate answer for more general geometry.
- In many numerical problems Nesterov is more efficient, but not always.
- The real answer is ... Nesterov may be more efficient than GD or not.

Outline

- Gradient descent and growth condition.
- State of the art on Nesterov scheme.
- New rates for Nesterov Schemes.
- Proofs coming from an ODE study.

Growth condition

A function F satisfies condition $\mathcal{L}(\gamma)$ if it exists $K > 0$ such that for all $x \in R^n$

$$d(x, X^*)^\gamma \leqslant K\left(F(x) - F(x^*)\right) \tag{6}$$

Theorem Garrigos al al.

- If F satisfies condition $\mathcal{L}(\gamma)$ with $\gamma > 2$ then

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{\alpha}{\alpha-2}}}\right) \tag{7}$$

- If F satisfies condition $\mathcal{L}(2)$ then it exists $a > 0$

$$F(x_n) - F(x^*) = O\left(e^{-an}\right) \tag{8}$$

Geometric convergence of GD with $\mathcal{L}(2)$

$$F(x_n) - F(x^*) \leqslant \frac{\|x_0 - x^*\|^2}{2hn} \text{ and } \|x - x^*\|^2 \leqslant K\left(F(x) - F(x^*)\right)$$

No memory algorithm $\Rightarrow \forall j \leqslant n$

$$F(x_n) - F(x^*) \leqslant \frac{\|x_{n-j} - x^*\|^2}{2hj} \leqslant \frac{K}{2hj}(F(x_{n-j}) - F(x^*))$$

If $\frac{K}{2hj} \leqslant \frac{1}{2} \iff j \geqslant \frac{K}{h}$,

$$F(x_n) - F(x^*) \leqslant \frac{F(x_{n-j}) - F(x^*)}{2}$$

Conclusion : The decay is geometric.

State of the art

- Nesterov Scheme for $h < \frac{1}{L}$, and $\alpha \geqslant 3$

$$x_{n+1} = x_n - h\nabla F\left(x_n + \frac{n}{n+\alpha}(x_n - x_{n-1})\right) \tag{9}$$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^2}\right) \tag{10}$$

- Chambolle, D (14) and Attouch Peypouquet (15):

$$\alpha > 3 \Rightarrow \text{convergence of } (x_n)_{n\geqslant 1} \text{ and } F(x_n) - F(x^*) = o\left(\frac{1}{n^2}\right)$$

- If $\alpha \leqslant 3$, Apidopoulos et al. and Attouch et al. (17)

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2\alpha}{3}}}\right) \tag{11}$$

Theorem Su Boyd Candès (15), Attouch Cabot (17)

If F satisfies $\mathcal{L}(2)$ and uniqueness of minimizer, then $\forall\alpha > 0$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2\alpha}{3}}}\right) \tag{12}$$

Another geometrical condition

Flatness condition

F satisfies condition $H(\gamma)$ if $\forall x \in \mathbb{R}^n$ and all $x^* \in X^*$

$$F(x) - F(x^*) \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^* \rangle \tag{13}$$

Flatness and growth properties

- If $(F - F^*)^{\frac{1}{\gamma}}$ is convex, then F satisfies $H1(\gamma)$.
- If F satisfies $H(\gamma)$ then it exists $K_2 > 0$ such that

$$F(x) - F(x^*) \leqslant K_2 d(x, X^*)^{\gamma} \tag{14}$$

- if $F(x) = \|x - x^*\|^r$, with $r > 1$, F satisfies $H1(\gamma)$ for all $\gamma \in [1, r]$ ... and $\mathcal{L}(p)$ for all $p \geqslant \gamma$.
- if F satisfies $\mathcal{L}(2)$ and ∇F is L-Lispchitz then F satisfies $H(1 + \frac{L}{2K_2})$.

Nesterov, flatness may improve convergence rate

Theorem : Apidopoulos et al. (18)

Let F be a differentiable convex function which gradient is $L-$Lipschitz

1. If F satisfies $H(\gamma)$, with $\gamma > 1$ and
   1. if $\alpha \leqslant 1 + \frac{2}{\gamma}$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2\gamma\alpha}{\gamma+2}}}\right) \tag{15}$$

   2. if $\alpha > 1 + \frac{2}{\gamma}$ and thus if $\alpha = 3$ then

$$F(x_n) - F(x^*) = o\left(\frac{1}{n^2}\right) \tag{16}$$

      and the sequence $(x_n)_{n\geqslant 1}$ converges.
2. If F satisfies $\mathcal{L}(2)$, the previous points apply for a $\gamma > 1$.

Theorem for sharp functions, Apidoupoulos et al. (18)

If F satisfies $\mathcal{L}(2)$, $H(\gamma)$ and has a unique minimizer x^* then $\forall \alpha > 0$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2\gamma\alpha}{\gamma+2}}}\right) \tag{17}$$

Comments

- For $\gamma = 1$ we recover the decay $O\left(\frac{1}{n^{\frac{2\alpha}{3}}}\right)$: Attouch and Cabot
- For quadratic functions, $\gamma = 2$ and thus we get $O\left(\frac{1}{n^\alpha}\right)$.
- Since ∇F is $L-$Lipschitz, F satisfies $H1(\gamma)$ for $\gamma > 1$ and thus $\frac{2\gamma\alpha}{\gamma+2} > \frac{2\alpha}{3}$.
- For $F(x) = \|Ax - y\|^2$ the decay is $O\left(\frac{1}{n^\alpha}\right)$.

Theorem for flat functions, Apidopoulos (18)

If F satisfies $H(\gamma)$ and $\mathcal{L}(\gamma)$ with $\gamma > 2$, if F has unique minimizer and if $\alpha > \frac{\gamma+2}{\gamma-2}$ then

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2\gamma}{\gamma-2}}}\right) \tag{18}$$

Gradient descent rate

If F satisfies $\mathcal{L}(\gamma)$ with $\gamma > 2$

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{\gamma}{\gamma-2}}}\right) \tag{19}$$

Discretization of an ODE, Su Boyd and Candès (15)

The scheme defined by

$$x_{n+1} = y_n - h\nabla F(y_n) \text{ with } y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}) \tag{20}$$

is a discretization of a solution of

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0 \tag{ODE}$$

With $\dot{x}(t_0) = 0$.
Move of a solid in a potential field with a vanishing viscosity $\frac{\alpha}{t}$.

Advantages of the discret setting

1. A simpler Lyapunov analysis, better insight
2. Optimality of bounds

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0 \tag{ODE}$$

Nesterov, Continuous

If F is convex and if $\alpha \geqslant 3$, the solution of (**??**) satisfies

$$F(x(t)) - F(x^*) = O\left(\frac{1}{t^2}\right) \tag{21}$$

$$x_{n+1} = y_n - h\nabla F(y_n) \text{ with } y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1})$$

Nesterov, Discret

If F is convex and if $\alpha \geqslant 3$, the sequence$(x_n)_{n\geqslant 1}$ satisfies

$$F(x_n) - F(x^*) = O\left(\frac{1}{n^2}\right) \tag{22}$$

Nesterov, Proof of the continuous theorem

We define

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \frac{1}{2}\left\|(\alpha - 1)(x(t) - x^*) + t\dot{x}(t)\right\|^2$$

Using (**??**) and the following convex inequality

$$F(x(t)) - F(x^*) \leqslant \langle x(t) - x^*, \nabla F(x(t))\rangle$$

we get

$$\mathcal{E}'(t) \leqslant (3 - \alpha)t(F(x(t) - F(x^*)) \tag{23}$$

1. If $\alpha \geqslant 3$, $\forall t \geqslant t_0$, $t^2(F(x(t)) - F(x^*)) \leqslant \mathcal{E}(t_0)$
2. If $\alpha > 3$, $\displaystyle\int_{t=t_0}^{+\infty}(\alpha - 3)t(F(x(t) - F(x^*)) \leqslant \mathcal{E}(t_0)$

Nesterov, Proof of the discret theorem

We define

$$\mathcal{E}_n = n^2(F(x_n) - F(x^*)) + \frac{1}{2h}\left\|(\alpha - 1)(x_n - x^*) + n(x_n - x_{n-1})\right\|^2$$

Using the definition of $(x_n)_{n\geqslant 1}$ and the following convex inequality

$$F(x_n) - F(x^*) \leqslant \langle x_n - x^*, \nabla F(x_n)\rangle$$

we get

$$\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant (3-\alpha)n(F(x_n) - F(x^*)) \tag{24}$$

1. If $\alpha \geqslant 3$, $\forall n \geqslant 1$, $n^2(F(x_n) - F(x^*)) \leqslant \mathcal{E}_1$
2. If $\alpha > 3$, $\sum_{n\geqslant 1}(\alpha - 3)n(F(x_n) - F(x^*)) \leqslant \mathcal{E}_1$

Nesterov, Proof of convergence rate

1. We define for $(p,\xi,\lambda)\in\mathbb{R}^3$

$$\mathcal{H}(t)=t^p(t^2(F(x(t))-F(x^*))+\frac{1}{2}\|(\lambda(x(t)-x^*)+t\dot{x}(t)\|^2+\frac{\xi}{2}\|x(t)-x^*\|^2)$$

2. We choose $(p,\xi,\lambda)\in\mathbb{R}^3$ depending on the hypotheses to ensure that $\mathcal{H}$ is bounded. $\mathcal{H}$ may not be non increasing.
3. We deduce there is $A\in\mathbb{R}$ such that

$$t^{2+p}(F(x(t))-F(x^*))\leqslant A-t^p\frac{\xi}{2}\|x(t)-x^*\|^2$$

4. If $\xi\geqslant 0$ then $F(x(t))-F(x^*)=O\left(\frac{1}{t^{p+2}}\right)$.
5. if $\xi\geqslant 0$ we must use conditions $\mathcal{L}(\gamma)$ to conclude.

Theorem Su, Boyd, Candès (15)

If F is convex, satisfies and $\alpha \geqslant 3$

$$F(x(t)) - F(x^*) = O\left(\frac{1}{t^2}\right) \tag{25}$$

Proof : $p = 0$, $\lambda = \alpha - 1, \xi = 0$

Theorem Aujol, D., Rondepierre (18)

If F is convex, satisfies $H(\gamma)$ and $\mathcal{L}(2)$, and has unique minimizer

$$F(x(t)) - F(x^*) = O\left(\frac{1}{t^{\frac{2\alpha\gamma}{\gamma+2}}}\right) \tag{26}$$

Proof : $p = \frac{2\alpha\gamma}{\gamma+2} - 2$, $\lambda = \frac{2\alpha}{\gamma+2}, \xi = \lambda(\lambda + 1 - \alpha)$.