Nesterov's Accelerated Gradient Descent on L-smooth convex function

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft: August 2, 2017
Last update : August 10, 2020

Overview

1. Nesterov's accelerated gradient descent (NAGD)
2. Proving NAGD converges at $\mathcal{O}\left(\frac{1}{k^2}\right)$
3. Summary

Gradient Descent (GD)

- Given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is convex and L-smooth [1] , the unconstrained convex minimization problem

$$\min_x f(x),$$

can be solved by Gradient Descent (GD) algorithm

- Starting with an initial point $x_0 \in \mathbb{R}^n$, GD iterates the update

$$x_{k+1} = x_k - t_k \nabla f(x_k).$$

If step size is sufficiently small ($t_k \leq \frac{2}{L}$), then $\{x_k\}_{k \in \mathbb{N}}$ converges to a stationary point of f. As f is convex, the sequence converges to the global minimizer x^* (if it exists).

- GD algorithm on such f has convergence rate $\mathcal{O}\left(\frac{1}{k}\right)$.

[1] f is L smooth if ∇f is L-Lipschitz. If f is not differentiable, ∇f is replaced by sub-gradient $g \in \partial f$.

Nesterov's Accelerated Gradient Descent(NAGD)

- On the same problem ($\min_{x \in \mathbb{R}^n} f(x)$, f is convex and L-smooth), NAGD iterates the following update scheme:

$$\begin{aligned}
\text{Gradient update} \quad y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) && (1)\\
\text{Extrapolation} \quad x_{k+1} &= (1-\gamma_k)y_{k+1} + \gamma_k y_k && (2)\\
\text{Extrapolation weight} \quad \gamma_k &= \frac{1-\lambda_k}{\lambda_{k+1}} && (3)\\
\text{Extrapolation weight} \quad \lambda_k &= \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2} && (4)
\end{aligned}$$

  with initial point $y_0 = x_0 \in \mathbb{R}^n$ and $\lambda_0 = 0$.

- Note that here fix stepsize is used: $t_k = \frac{1}{L}$.

NAGD has convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$

Theorem. If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and convex, the sequences $f(y_k)$ produced by the NAGD algorithm converges to the optimal value $f^* = f(x^*)$ at the rate $\mathcal{O}\left(\frac{1}{k^2}\right)$ as

$$f(y_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{k^2}.$$

Note.

- It can be proved that the convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$ is optimal. i.e., no 1st-order algorithm can perform better than NAGD in terms of convergence rate. All 1st-order algorithm can only be at most as good as NAGD.
- If f is nonconvex, the sequence $f(y_k)$ produced by the NAGD algorithm will converge to the closest stationary point with the same convergence rate.

NAGD converges at $\mathcal{O}\left(\frac{1}{k^2}\right)$ - the proof ... 1/11

- Tool 1 (make use of the fact that f is convex) For all $x, y \in \mathrm{dom} f$
$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$
It gives
$$-f(y) \leq -f(x) + \nabla f(x)^T (x - y). \tag{5}$$

- Tool 2. (make use of the fact that f is L-smooth) For all $a, b \in \mathrm{dom} f$
$$f(a) - f(b) \leq \nabla f(b)^T (a - b) + \frac{L}{2} \|a - b\|_2^2.$$
Put $a = x - \frac{1}{L} \nabla f(x)$ and $b = x$,
$$\begin{aligned} f\left(x - \frac{1}{L}\nabla f(x)\right) - f(x) &\leq -\frac{1}{L}\|\nabla f(x)\|_2^2 + \frac{1}{2L}\|\nabla f(x)\|_2^2 \\ &= -\frac{1}{2L}\|\nabla f(x)\|_2^2. \end{aligned} \tag{6}$$

- (5) + (6), the $-f(x)$ term is canceled:
$$f\left(x - \frac{1}{L}\nabla f(x)\right) - f(y) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2 + \nabla f(x)^T (x - y). \tag{7}$$

The proof ... 2/11

- Restate (7)

$$f\left(x - \frac{1}{L}\nabla f(x)\right) - f(y) \le \frac{-1}{2L}\|\nabla f(x)\|_2^2 + \nabla f(x)^T(x - y).$$

- Put $x = x_k$, $y = x^*$ in (7)

$$f\left(x_k - \frac{1}{L}\nabla f(x_k)\right) - f^* \le \frac{-1}{2L}\|\nabla f(x_k)\|_2^2 + \nabla f(x_k)^T(x_k - x^*) \tag{8}$$

- put $x = x_k$, $y = y_k$ in (7)

$$f\left(x_k - \frac{1}{L}\nabla f(x_k)\right) - f(y_k) \le \frac{-1}{2L}\|\nabla f(x_k)\|_2^2 + \nabla f(x_k)^T(x_k - y_k) \tag{9}$$

- High-level overview of the proof: at this stage we established two inequalities (8,9) that link $f(y_{k+1}), f(y_k)$ and f^*. We see that the two inequalities contain the gradient term $\nabla f(x_k)$. The convergence result does not have the gradient term $\nabla f(x_k)$, so what we do next is to try to remove such term in the inequalities.

The proof ... 3/11

- To simplify notation, define $\delta_k := f(y_k) - f^*$, then

$$\begin{aligned}
f\Big(x_k - \frac{1}{L}\nabla f(x_k)\Big) &\overset{(1)}{=} f(y_{k+1}) && (10)\\
f\Big(x_k - \frac{1}{L}\nabla f(x_k)\Big) - f^* &= \delta_{k+1} && (11)\\
f\Big(x_k - \frac{1}{L}\nabla f(x_k)\Big) - f(y_k) &= f\Big(x_k - \frac{1}{L}\nabla f(x_k)\Big) - f^* - \Big(f(y_k) - f^*\Big) \\
&= \delta_{k+1} - \delta_k && (12)\\
\nabla f(x_k) &\overset{(1)}{=} -L(y_{k+1} - x_k) && (13)\\
\|\nabla f(x_k)\|_2^2 &\overset{(13)}{=} L^2\|y_{k+1} - x_k\|_2^2 && (14)
\end{aligned}$$

- Put (11,13,14) into (8)

$$\delta_{k+1} \leq -\frac{L}{2}\|y_{k+1} - x_k\|_2^2 - L(y_{k+1} - x_k)^T(x_k - x^*). \quad (15)$$

- Put (12,13,14) into (9)

$$\delta_{k+1} - \delta_k \leq -\frac{L}{2}\|y_{k+1} - x_k\|_2^2 - L(y_{k+1} - x_k)^T(x_k - y_k). \quad (16)$$

The proof ... 4/11 (Things start to go crazy here)

- Tricky step : consider $(15) + (\lambda_k - 1)(16)$ so as to combine (15) and (16). On the left hand side we have

$$\delta_{k+1} + (\lambda_k - 1)(\delta_{k+1} - \delta_k) \quad = \quad \lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k.$$

- On the right hand side we have

$$\begin{aligned}
& -\frac{L}{2}\|y_{k+1} - x_k\|_2^2 - L(y_{k+1} - x_k)^T(x_k - x^*) \\
& -(\lambda_k - 1)\left(\frac{L}{2}\|y_{k+1} - x_k\|_2^2 - L(y_{k+1} - x_k)^T(x_k - y_k)\right) \\
= \quad & -\frac{\lambda_k L}{2}\|y_{k+1} - x_k\|_2^2 - L(y_{k+1} - x_k)^T\Big(x_k - x^* + (\lambda_k - 1)(x_k - y_k)\Big) \\
= \quad & -\frac{\lambda_k L}{2}\|y_{k+1} - x_k\|_2^2 - L(y_{k+1} - x_k)^T\Big(\lambda_k x_k - (\lambda_k - 1)y_k - x^*\Big).
\end{aligned}$$

- Hence we have

$$\lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k \quad \leq \quad \begin{aligned} & -\frac{\lambda_k L}{2}\|y_{k+1} - x_k\|_2^2 \\ & -L(y_{k+1} - x_k)^T\Big(\lambda_k x_k - (\lambda_k - 1)y_k - x^*\Big). \end{aligned} \tag{17}$$

The proof ... 5/11

Multiply (17) inequality with λ_k:

$$\lambda_k^2\delta_{k+1} - \lambda_k(\lambda_k-1)\delta_k \quad \leq \quad -\frac{\lambda_k^2 L}{2}\|y_{k+1}-x_k\|_2^2 - \lambda_k L(y_{k+1}-x_k)^T\Big(\lambda_k x_k - (\lambda_k-1)y_k - x^*\Big)$$

From (4) $\lambda_k = \frac{1}{2}\Big(1+\sqrt{1+4\lambda_{k-1}^2}\Big)$, we get

$$(2\lambda_k-1)^2 = 1+4\lambda_{k-1}^2 \iff \lambda_{k-1}^2 = \lambda_k(\lambda_k-1).$$

Put $\lambda_{k-1}^2 = \lambda_k(\lambda_k-1)$ in the equation above

$$\begin{aligned}
\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \quad &\leq \quad -\frac{\lambda_k^2 L}{2}\|y_{k+1}-x_k\|_2^2 \\
&\qquad - \lambda_k L(y_{k+1}-x_k)^T\big(\lambda_k x_k - (\lambda_k-1)y_k - x^*\big) \\
\text{[re-arrange RHS]} \quad &= \quad -\frac{L}{2}\Big(\lambda_k^2\|y_{k+1}-x_k\|_2^2 \\
&\qquad + 2\lambda_k(y_{k+1}-x_k)^T(\lambda_k x_k - (\lambda_k-1)y_k - x^*)\Big). \qquad (18)
\end{aligned}$$

The proof ... 6/11

Super tricky step: the expression

$$\lambda_k^2\|y_{k+1} - x_k\|_2^2 + 2\lambda_k(y_{k+1} - x_k)^T(\lambda_k x_k - (\lambda_k - 1)y_k - x^*).$$

is equivalent to

$$\|\lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*\|_2^2 + \|\lambda_k x_k - (\lambda_k - 1)y_k - x^*\|_2^2.$$

Why:

$$\lambda^2(y-x)^2 + 2\lambda(y-x)(\lambda x - (\lambda-1)z - w) = (\lambda y - (\lambda-1)z - w)^2 + (\lambda x - (\lambda-1)z - w)^2.$$

Apply to (18)

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \quad \leq \quad -\frac{L}{2}\Big(\|\lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*\|_2^2 + \|\lambda_k x_k - (\lambda_k - 1)y_k - x^*\|_2^2\Big). \tag{19}$$

The proof ... 7/11

We have the following equality:

$$\lambda_k x_k - (\lambda_k - 1)y_k = (1 - \lambda_{k-1})y_{k-1} + \lambda_{k-1}y_k.$$

Proof: By (3) $\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}$, $\gamma_k \lambda_{k+1} = 1 - \lambda_k$.
By (2) $x_{k+1} = (1-\gamma_k)y_{k+1} + \gamma_k y_k$ get $x_{k+1} = y_{k+1} + \gamma_k(y_k - y_{k+1})$, multiply with λ_{k+1} gives $\lambda_{k+1}x_{k+1} = \lambda_{k+1}y_{k+1} + \lambda_{k+1}\gamma_k(y_k - y_{k+1}) = \lambda_{k+1}y_{k+1} + (1-\lambda_k)(y_k - y_{k+1})$, rearrange this get
$\lambda_{k+1}x_{k+1} - \lambda_{k+1}y_{k+1} = (1-\lambda_k)(y_k - y_{k+1})$, add y_{k+1} on both side get
$\lambda_{k+1}x_{k+1} - (\lambda_{k+1} - 1)y_{k+1} = (1-\lambda_k)y_k + \lambda_k y_{k+1}$.

---

hence (19) becomes

$$\lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k \leq -\frac{L}{2}\Big(\|\lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*\|_2^2 + \|(1-\lambda_{k-1})y_{k-1} + \lambda_{k-1}y_k - x^*\|_2^2 \Big). \tag{20}$$

The proof ... 8/11

Rearrange (20) that the two terms in right hand side have similar form

$$\begin{aligned}\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \quad \leq \quad & -\frac{L}{2}\Big(\|\lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*\|_2^2 \\ & \quad + \|\lambda_{k-1}y_k - (\lambda_{k-1} - 1)y_{k-1} - x^*\|_2^2\Big).\end{aligned} \tag{21}$$

Let $u_k = \lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*$ so
$\lambda_{k-1}y_k - (\lambda_{k-1} - 1)y_{k-1} - x^* = u_{k-1}$ and (21) becomes

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq -\frac{L}{2}\Big(\|u_k\|_2^2 - \|u_{k-1}\|_2^2\Big).$$

which forms a telescoping series !

The proof ... 9/11

$$
\begin{array}{ll}
k = 1 & \lambda_1^2\delta_2 - \lambda_0^2\delta_1 \le -\frac{L}{2}\left(\|u_1\|_2^2 - \|u_0\|_2^2\right) \\
k = 2 & \lambda_2^2\delta_3 - \lambda_1^2\delta_2 \le -\frac{L}{2}\left(\|u_2\|_2^2 - \|u_1\|_2^2\right) \\
\vdots & \vdots \\
k = K-1 & \lambda_{K-1}^2\delta_K - \lambda_{K-2}^2\delta_{K-1} \le -\frac{L}{2}\left(\|u_{K-1}\|_2^2 - \|u_{K-2}\|_2^2\right) \\
\text{Sum all} & \lambda_{K-1}^2\delta_K - \lambda_0^2\delta_1 \le -\frac{L}{2}\left(\|u_{K-1}\|_2^2 - \|u_0\|_2^2\right) \\
\text{rearrange} & \lambda_{K-1}^2\delta_K - \lambda_0^2\delta_1 \le \frac{L}{2}\left(\|u_0\|_2^2 - \|u_{K-1}\|_2^2\right)
\end{array}
$$

By definition, $\lambda_0 = 0$, $u_0 = \lambda_0 y_1 - (\lambda_0 - 1)y_0 - x^* = y_0 - x^*$, and $y_0 = x_0$:

$$\lambda_{K-1}^2\delta_K \le \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|u_{K-1}\|_2^2\right).$$

As $\|u_{K-1}\|_2^2 \ge 0$

$$\delta_K \le \frac{L\|x_0 - x^*\|_2^2}{2\lambda_{K-1}^2}.$$

The proof ... 10/11

Lemma. $\lambda_{k-1} \geq \frac{k}{2}$.

Proof by induction. For $k = 0$ it is trivial ($0 \geq 0/2$).

When $k = 1$, by definition $\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$, so $k_1 = 1 > \frac{1}{2} = \left[\frac{k}{2}\right]_{k=1}$

(Induction hypothesis). Assume $\lambda_{n-1} \geq \frac{n}{2}$.

When $k = n$,

$$\begin{aligned}
\lambda_n &= \frac{1+\sqrt{1+4\lambda_{n-1}^2}}{2} \\
\text{[Induction hypothesis]} \quad &\geq \frac{1+\sqrt{1+4\left(\frac{n}{2}\right)^2}}{2} \\
&= \frac{1+\sqrt{1+n^2}}{2} \\
&> \frac{1+\sqrt{n^2}}{2} = \frac{1+n}{2}. \quad \square
\end{aligned}$$

The proof ... 11/11

With $\lambda_{k-1} \geq \frac{k}{2}$, so $\frac{1}{\lambda_{k-1}^2} \leq \frac{4}{k^2}$ and

$$\delta_K \leq \frac{L\|x_0 - x^*\|_2^2}{2\lambda_{K-1}^2}$$

becomes

$$f_{y_k} - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{k^2}.$$

where $f_{y_k} - f^* = \delta_k$ (by definition).

The (crazy, complicated, highly-involved, non-intuitive) proof is now completed. $\square$

Last page - summary

For unconstrained convex smooth problem $\min_{x \in \mathbb{R}^n} f(x)$, with $f : \mathbb{R}^n \to \mathbb{R}$ being L-smooth and convex, the NAGD algorithm starts with initial point $x_0 = y_0 \in \mathbb{R}^n$ and $\lambda_0 = 0$ and iterates the following:

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \qquad x_{k+1} = (1-\gamma_k)y_{k+1} + \gamma_k y_k$$

$$\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}} \qquad \lambda_k = \frac{1}{2}\left(1 + \sqrt{1+4\lambda_{k-1}^2}\right),$$

the sequences $f(y_k)$ produced will converges to the optimal value f^* at order of $\mathcal{O}\left(\frac{1}{k^2}\right)$ as

$$f(y_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{k^2}.$$

End of document