A 2n-order ODE of Nesterov's accelerated gradient

$$\ddot{\boldsymbol{X}}_t + \frac{3}{t}\dot{\boldsymbol{X}}_t + \nabla f(\boldsymbol{X}_t) = 0.$$

Andersen Ang

Dept. Combinatorics & Optimization, University of Waterloo, Canada

ms$\boldsymbol{x}$ang@uwaterloo.ca, where $\boldsymbol{x} = \lfloor \pi \rfloor$ Homepage: angms.science

First draft: November 15, 2020 Last update: December 9, 2022

Nesterov's accelerated gradient (NAG)

- Let α_k be stepsize and β_k be extrapolation parameter,

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{y}_k - \alpha_k \nabla f(\boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1} &= \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k). \end{aligned} \tag{NAG}$$

- On can pick β_k as[1]

$$\beta_k = \frac{k}{k+3} \quad \text{or} \quad \beta_k = \frac{k-1}{k+2}.$$

- Theorem: if f is convex and L-smooth, with stepsize $\alpha_k < \frac{2}{L}$, NAG has convergence rate

$$f(\boldsymbol{x}_k) - f^* \leq \frac{\text{constant}}{(k+1)^2} = \mathcal{O}\left(\frac{1}{k^2}\right),$$

where $f^* = \min f$.

[1] This is not the one proposed by Nesterov in 1983 but it satisfies the Paul Tseng's rule, see (15) in "On Accelerated Proximal Gradient Methods for Convex-Concave Optimization".

Su-Boyd-Candes ODE

- It was shown[2] NAG associates to

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{y}_k - \alpha_k \nabla f(\boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1} &= \boldsymbol{x}_{k+1} + \frac{1}{k+3}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \end{aligned} \iff \ddot{\boldsymbol{X}}_t + \frac{3}{t}\dot{\boldsymbol{X}}_t + \nabla f(\boldsymbol{X}_t) = \mathbf{0}. \quad \text{(SBC)}$$

  Under the specific setting $t = k\sqrt{\alpha}$ with $\beta_k = \dfrac{1}{k+3}$

- Notation
  - $\boldsymbol{x}_k \in \mathbb{R}^n$ is the optimization variable on discrete time k.
  - $\boldsymbol{X}_t$ is the variable on the continuous time t.

[2] W. Su, S. Boyd, and E. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights" in NIPS2014 and JMLR2016

Prerequisite for deriving the ODE

1. (2nd-order) Taylor series
   Given a function u that is twice-differentiable, the Taylor series at x_0 with a change Δ is

$$u(x_0 + \Delta) = u(x_0) + \Delta \frac{\partial u}{\partial x}\Big|_{x=x_0} + \frac{\Delta^2}{2!}\frac{\partial^2 u}{\partial^2 x}\Big|_{x=x_0} + o(\Delta).$$

   where $o(\Delta) =$ higher-order terms of Δ.

2. Time-derivative notation: $\dot{\square} := \frac{d}{dt}\square, \quad \ddot{\square} := \frac{d^2}{dt^2}\square$

3. $\lim_{\alpha \to 0} \sqrt{\alpha} = 0.$

Derive the ODE: forming finite differences

- For NAG with constant stepsize: $\alpha_k = \alpha$

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{y}_k - \alpha \nabla f(\boldsymbol{y}_k), & (1) \\ \boldsymbol{y}_{k+1} &= \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k). & (2) \end{aligned}$$

- From (2), consider at $k-1$

$$\boldsymbol{y}_{k} = \boldsymbol{x}_{k} + \beta_{k-1}(\boldsymbol{x}_{k} - \boldsymbol{x}_{k-1}). \tag{3}$$

- Put (3) into (1)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_{k} + \beta_{k-1}(\boldsymbol{x}_{k} - \boldsymbol{x}_{k-1}) - \alpha \nabla f(\boldsymbol{y}_k).$$

  Rearrange

$$\underbrace{\boldsymbol{x}_{k+1} - \boldsymbol{x}_{k}}_{\text{finite difference}} = \beta_{k-1}(\underbrace{\boldsymbol{x}_{k} - \boldsymbol{x}_{k-1}}_{\text{finite difference}}) - \alpha \nabla f(\boldsymbol{y}_k). \tag{4}$$

- What we did: combine two steps to 1 and obtain finite difference terms. Now our goal is to derive an ODE from (4).

Derive the ODE: discretization $t = k\sqrt{\alpha}$

- To derive an $\underbrace{\text{differential equation}}_{\text{in continuous time}}$ from $\underbrace{\text{finite difference Eq. (4)}}_{\text{in discrete time}}$, we need to link the times t and k.
- Naturally we let $t = kh$ where $h = \alpha$ is the discretization stepsize. But the proof does not work.
- Instead, pick $h = \sqrt{\alpha}$, i.e., choose the discretization stepsize as the square-root of gradient stepsize, we get

$$t = k\sqrt{\alpha} \quad : \quad \text{continuous time} \;=\; \text{discrete iteration} \times \text{sqrt gradient stepsize.}$$

- Now we have approximation

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{X}_{k\sqrt{\alpha}} &&= \boldsymbol{x}_k + o(\sqrt{\alpha}) \\ \boldsymbol{X}_{t+\sqrt{\alpha}} &= \boldsymbol{X}_{(k+1)\sqrt{\alpha}} &&= \boldsymbol{x}_{k+1} + o(\sqrt{\alpha}) \\ \boldsymbol{X}_{t-\sqrt{\alpha}} &= \boldsymbol{X}_{(k-1)\sqrt{\alpha}} &&= \boldsymbol{x}_{k-1} + o(\sqrt{\alpha}) \end{aligned} \qquad \begin{aligned} \boldsymbol{x}_{k+1} - \boldsymbol{x}_k &= \boldsymbol{X}_{t+\sqrt{\alpha}} - \boldsymbol{X}_t + o(\sqrt{\alpha}) \\ \boldsymbol{x}_k - \boldsymbol{x}_{k-1} &= \boldsymbol{X}_t - \boldsymbol{X}_{t-\sqrt{\alpha}} + o(\sqrt{\alpha}) \end{aligned}$$

- $\boldsymbol{X}_{t+\sqrt{\alpha}}$ is in the form $u(x_0 + \Delta) \implies$ use Taylor's series!

Derive the ODE: Taylor's series $u(x_0+\Delta) = u(x_0) + \Delta \frac{\partial u}{\partial x}\Big|_{x=x_0} + \frac{\Delta^2}{2!}\frac{\partial^2 u}{\partial^2 x}\Big|_{x=x_0} + o(\Delta)$

- On $\boldsymbol{X}(t+\sqrt{\alpha})$: let $u = \boldsymbol{X}$, $x_0 = t$, $\Delta = \sqrt{\alpha}$ and $\frac{\partial u}{\partial x} = \frac{\partial \boldsymbol{X}}{\partial t} = \dot{\boldsymbol{X}}$, then

$$\boldsymbol{X}(t+\sqrt{\alpha}) = \boldsymbol{X}(t) + \sqrt{\alpha}\dot{\boldsymbol{X}}(t) + \frac{\alpha}{2}\ddot{\boldsymbol{X}}(t) + o(\sqrt{\alpha}).$$

- On $\boldsymbol{X}(t-\sqrt{\alpha})$: let $u = \boldsymbol{X}$, $x_0 = t$, $\Delta = -\sqrt{\alpha}$ and $\frac{\partial u}{\partial x} = \frac{\partial \boldsymbol{X}}{\partial t} = \dot{\boldsymbol{X}}$, then

$$\boldsymbol{X}(t-\sqrt{\alpha}) = \boldsymbol{X}(t) - \sqrt{\alpha}\dot{\boldsymbol{X}}(t) + \frac{\alpha}{2}\ddot{\boldsymbol{X}}(t) + o(\sqrt{\alpha}).$$

  Note: you don't care about the sign in $o(\sqrt{\alpha})$ here.

- We have

$$\begin{aligned}
\boldsymbol{x}_{k+1} - \boldsymbol{x}_k &= \boldsymbol{X}(t+\sqrt{\alpha}) - \boldsymbol{X}(t) + o(\sqrt{\alpha}) = \sqrt{\alpha}\dot{\boldsymbol{X}} + \frac{\alpha}{2}\ddot{\boldsymbol{X}} + o(\sqrt{\alpha}) \\
\boldsymbol{x}_k - \boldsymbol{x}_{k-1} &= \boldsymbol{X}(t) - \boldsymbol{X}(t-\sqrt{\alpha}) + o(\sqrt{\alpha}) = \sqrt{\alpha}\dot{\boldsymbol{X}} - \frac{\alpha}{2}\ddot{\boldsymbol{X}} + o(\sqrt{\alpha})
\end{aligned}$$

  Note: be careful of the sign.

Derive the ODE: almost there

- Recall finite difference equation (4) is

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = \beta_{k-1}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \alpha \nabla f(\boldsymbol{y}_k). \tag{4}$$

- Now (4) becomes

$$\sqrt{\alpha}\dot{\boldsymbol{X}}_t + \frac{\alpha}{2}\ddot{\boldsymbol{X}}_t + o(\sqrt{\alpha}) = \beta_{k-1}\left(\sqrt{\alpha}\dot{\boldsymbol{X}}_t - \frac{\alpha}{2}\ddot{\boldsymbol{X}}_t + o(\sqrt{\alpha})\right) - \alpha \nabla f(\boldsymbol{y}_k).$$

- Rearrange

$$\frac{\alpha}{2}(1+\beta_{k-1})\ddot{\boldsymbol{X}}_t + \sqrt{\alpha}(1-\beta_{k-1})\dot{\boldsymbol{X}}_t + \alpha \nabla f(\boldsymbol{y}_k) + o(\sqrt{\alpha}) = 0.$$

- For $\boldsymbol{y}_k$, as $\boldsymbol{x}_k = \boldsymbol{y}_k$ in the long run, we can take $\boldsymbol{y}_k = \boldsymbol{Y}_t + o(\sqrt{\alpha})$ and $\boldsymbol{Y}_t = \boldsymbol{X}_t$.

$$\frac{\alpha}{2}(1+\beta_{k-1})\ddot{\boldsymbol{X}}_t + \sqrt{\alpha}(1-\beta_{k-1})\dot{\boldsymbol{X}}_t + \alpha \nabla f(\boldsymbol{X}_t) + o(\sqrt{\alpha}) = 0. \tag{$*$}$$

The β_k

- There are in fact infinitely many choice of β, as long as it satisfies[3] $\dfrac{1-\beta_{k+1}}{\beta_{k+1}^2} \leq \dfrac{1}{\beta_k^2}$.
- What we want: taking $\lim\limits_{\alpha \to 0}$ will not remove $\ddot{\boldsymbol{X}}_t$ nor blow up the ODE.
- Try $\beta_k = \dfrac{k}{k+3}$. Now $\beta_{k-1} = \dfrac{k-1}{k+2} = 1 - \dfrac{-3}{k+2} \overset{k \gg 2}{\approx} 1 - \dfrac{3}{k} \overset{k=\frac{t}{\sqrt{\alpha}}}{=} 1 - \dfrac{3\sqrt{\alpha}}{t}$.
- $\beta_k = \dfrac{k}{k+3}$ satisfies $\dfrac{1-\beta_{k+1}}{\beta_{k+1}^2} \leq \dfrac{1}{\beta_k^2}$
  - For $k > 0$ we have $k+3 > k$ and hence $\dfrac{1}{k+3} < \dfrac{1}{k}$, which implies $\dfrac{k}{k+3} \leq \dfrac{k}{k} = 1$. Thus, $\beta_k = \dfrac{k}{k+3} \leq 1$.
  - L'Hopital's rule $\lim\limits_{k \to \infty} \dfrac{k}{k+3} = \lim\limits_{k \to \infty} \dfrac{\frac{d}{dk} k}{\frac{d}{dk}(k+3)} = \lim\limits_{k \to \infty} \dfrac{1}{1} = 1$. Thus $\beta_k = \dfrac{k}{k+3} \leq 1$ and is approaching to 1.
  - It is not hard to see $\beta_k > 0$ and thus $1 - \beta_{k+1} \overset{\beta_k \in [0,1]}{\leq} 1$
  - It is not hard to see β_k is an increasing sequence and thus $\dfrac{1}{\beta_{k+1}^2} \leq \dfrac{1}{\beta_k^2}$.
  - Multiply $1 - \beta_{k+1} \leq 1$ and $\dfrac{1}{\beta_{k+1}^2} \leq \dfrac{1}{\beta_k^2}$ we have $\dfrac{1-\beta_{k+1}}{\beta_{k+1}^2} \leq \dfrac{1}{\beta_k^2}$.

[3] Paul Tseng, "On Accelerated Proximal Gradient Methods for Convex-Concave Optimization".

Finishing the derivation with $\beta_{k-1} = 1 - \dfrac{3\sqrt{\alpha}}{t}$

- Now $(*)$
$$\frac{\alpha}{2}(1+\beta_{k-1})\ddot{\boldsymbol{X}}_t + \sqrt{\alpha}(1-\beta_{k-1})\dot{\boldsymbol{X}}_t + \alpha\nabla f(\boldsymbol{X}_t) + o(\sqrt{\alpha}) = 0 \qquad (*)$$
becomes
$$\frac{\alpha}{2}\left(2 - \frac{3\sqrt{\alpha}}{t}\right)\ddot{\boldsymbol{X}}_t + \sqrt{\alpha}\left(\frac{3\sqrt{\alpha}}{t}\right)\dot{\boldsymbol{X}}_t + \alpha\nabla f(\boldsymbol{X}_t) + o(\sqrt{\alpha}) = 0.$$

- Divide the whole equation by α, and note that $o(\sqrt{\alpha})$ contains terms with cubic power or higher in $\sqrt{\alpha}$ and hence they have α,
$$\frac{1}{2}\left(2 - \frac{3\sqrt{\alpha}}{t}\right)\ddot{\boldsymbol{X}}_t + \frac{3}{t}\dot{\boldsymbol{X}}_t + \nabla f(\boldsymbol{X}_t) + o(\sqrt{\alpha}) = 0.$$

- Take $\lim\limits_{\alpha\to 0}$ gives $\ddot{\boldsymbol{X}}_t + \dfrac{3}{t}\dot{\boldsymbol{X}}_t + \nabla f(\boldsymbol{X}_t) = 0.$

Why $h = \alpha$ does not work

- Consider instead of using $h = \sqrt{\alpha}$, pick $h = \alpha$. Then we have

$$\frac{\alpha^2}{2}\left(2 - \frac{3\alpha}{t}\right)\ddot{\boldsymbol{X}}_t + \alpha\left(\frac{3\alpha}{t}\right)\dot{\boldsymbol{X}}_t + \alpha\nabla f(\boldsymbol{X}_t) + o(\alpha) = 0.$$

Take $\lim\limits_{\alpha \to 0}$ makes the whole equation disappear.

- If we divide the whole equation by α^2, we have

$$\frac{1}{2}\left(2 - \frac{3\alpha}{t}\right)\ddot{\boldsymbol{X}}_t + \left(\frac{3}{t}\right)\dot{\boldsymbol{X}}_t + \frac{1}{\alpha}\nabla f(\boldsymbol{X}_t) + o(\alpha) = 0.$$

Take $\lim\limits_{\alpha \to 0}$ will blow up the gradient term.

- In fact it is the term $\alpha\nabla f$ in the very beginning determined to use $t = k\sqrt{\alpha}$.

Last page - summary

- Nesterov's accelerated gradient (NAG)

$$\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \alpha_k \nabla f(\boldsymbol{y}_k), \quad \boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k),$$

- Under $k = \dfrac{t}{\sqrt{\alpha}}$ with $\beta_k = \dfrac{k}{k+3}$, NAG is associated with the ODE

$$\ddot{\boldsymbol{X}}_t + \frac{3}{t}\dot{\boldsymbol{X}}_t + \nabla f(\boldsymbol{X}_t) = 0.$$

- Standard ODE theory (not discussed here) gives

$$f\Big(\boldsymbol{X}_t\Big) - f^* \leq \frac{\text{constant}}{t^2} = \mathcal{O}\Big(\frac{1}{t^2}\Big).$$

  As ODE $\iff$ NAG, this partially explains

$$f(\boldsymbol{x}_k) - f^* = \mathcal{O}\Big(\frac{1}{k^2}\Big).$$

End of document