# An 2n-order ODE dynamics that corresponds to Nesterov's accelerated gradient

Andersen Ang

Department of Combinatorics and Optimization,
University of Waterloo, Waterloo, Canada

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

First draft: November 15, 2020
Last update: March 15, 2021

# Nesterov's accelerated gradient

- Nesterov's accelerated gradient (NAG)

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

  where $\alpha_k$ is the stepsize and $\beta_k$ is the extrapolation parameter.
- On can pick $\beta$ as follows[1] for $\beta$, one can choose

$$\beta_k = \frac{k}{k+3}.$$

- Theorem of NAG: if $f$ is convex and $L$-smooth, picking stepsize $\alpha_k = \frac{1}{L}$, NAG has the convergence rate as

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

  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".

# Nesterov's accelerated gradient

- It can be shown that[2] NAG associates to the following 2nd-order ODE dynamics

$$\ddot{\mathbf{X}} + \frac{3}{t}\dot{\mathbf{X}} + \nabla f(\mathbf{X}) = \mathbf{0}. \tag{SBC}$$

  This document: show the derivation of this ODE.

- Notation
  - $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable.
  - $\mathbf{x}_k$ is the variable on discrete time $k$.
  - $\mathbf{X}(t)$ is the variable on the continuous time $t$.
  - $t$ may be omitted if it is clear from the context.

[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

- (2nd-order) Taylor series expansion
  Given a function $u$, the Taylor series expansion around a point $x_0$ with a small change $\Delta$ 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 $o(\Delta)$ holds the higher-order terms of $\Delta$.

- Time-derivative notation: $\dot{\square} := \frac{d}{dt}\square(t)$, $\ddot{\square} := \frac{d^2}{dt^2}\square(t)$

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

# Derive the ODE: forming finite difference terms

- Consider NAG with constant stepsize: $\alpha_k = \alpha$ and

$$\mathbf{x}_{k+1} = \mathbf{y}_k - \alpha \nabla f(\mathbf{y}_k), \tag{1}$$
$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \beta_k(\mathbf{x}_{k+1} - \mathbf{x}_k). \tag{2}$$

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

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

- Put (3) into (1)

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

  Rearrange

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

- Review what we did:
  - We canceled the $\mathbf{y}$ by combining the two steps in NAG, and
  - We rearrange to get the finite difference terms $\mathbf{x}_{k+1} - \mathbf{x}_k$.
- Now our goal is to derive an ODE from (4).

# Derive the ODE: discretization (1/2)

- Equation (4) lives in discrete time, and ODE lives in continuous time.

- To derive an ODE in the continuous time from an equation in discrete time, we need to link the time units, one way is to let $t = kh$ where $h$ is the discretization stepsize.

- If we pick $h = \sqrt{\alpha}$, i.e., choosing the discretization stepsize as the square-root of the stepsize of the gradient update, we get

$$k = \frac{t}{\sqrt{\alpha}} \ : \ \text{discrete iteration} = \frac{\text{continuous time}}{\text{stepsize}}. \tag{5}$$

- An immediate question is that why choose $h = \sqrt{\alpha}$ but not $\alpha$ in (5). Answer: because the derivation does not work with $h = \alpha$.

# Derive the ODE: discretization (2/2)

- Based on $h = \sqrt{\alpha}$ and (5) that $k = \dfrac{t}{h} = \dfrac{t}{\sqrt{\alpha}}$, then

$$\begin{array}{rcccl} \mathbf{X}(t) & = & \mathbf{X}(k\sqrt{\alpha}) & \approx & \mathbf{x}_k \\ \mathbf{X}(t+\sqrt{\alpha}) & = & \mathbf{X}\big((k+1)\sqrt{\alpha}\big) & \approx & \mathbf{x}_{k+1} \\ \mathbf{X}(t-\sqrt{\alpha}) & = & \mathbf{X}\big((k-1)\sqrt{\alpha}\big) & \approx & \mathbf{x}_{k-1} \end{array}$$

  and the "$\approx$" becomes "$=$" if we take $\lim\limits_{\alpha \to 0}$.

- Using these approximation, the finite difference terms now become

$$\begin{array}{rcl} \mathbf{x}_{k+1} - \mathbf{x}_k & \approx & \mathbf{X}(t+\sqrt{\alpha}) - \mathbf{X}(t) \\ \mathbf{x}_k - \mathbf{x}_{k-1} & \approx & \mathbf{X}(t) - \mathbf{X}(t-\sqrt{\alpha}) \end{array}$$

- Now the term $\mathbf{X}(t+\sqrt{\alpha})$ has the form $u(x_0 + \Delta)$, we can use Taylor's approximation on it.

# Derive the ODE: Taylor's approximation (1/2)

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

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

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

- For $\mathbf{X}(t-\sqrt{\alpha})$:
  Let $u = \mathbf{X}$, $x_0 = t$, $\Delta = -\sqrt{\alpha}$ and $\frac{\partial u}{\partial x} = \frac{\partial \mathbf{X}}{\partial t} = \dot{\mathbf{X}}$, then

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

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

# Derive the ODE: Taylor's approximation (2/2)

- Based on the Taylor's approximation,

$$
\begin{aligned}
\mathbf{X}(t+\sqrt{\alpha}) &= \mathbf{X}(t)+\sqrt{\alpha}\dot{\mathbf{X}}(t)+\frac{\alpha}{2}\ddot{\mathbf{X}}(t)+o(\sqrt{\alpha}) \\
\mathbf{X}(t-\sqrt{\alpha}) &= \mathbf{X}(t)-\sqrt{\alpha}\dot{\mathbf{X}}(t)+\frac{\alpha}{2}\ddot{\mathbf{X}}(t)+o(\sqrt{\alpha}).
\end{aligned}
$$

- We have

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

which becomes

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

Note: be careful of the sign.

# Derive the ODE: almost there

- Now (4) becomes

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

- Rearrange

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

- For $\mathbf{y}_k$, as $\mathbf{x}_k = \mathbf{y}_k$ in the long run, we can take $\mathbf{y}_k = \mathbf{X}(t)$.

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

- What next: plug-in $\beta_{k-1}$ and take $\lim\limits_{\alpha\to 0}$.

# The $\beta_k$

- There are in fact infinitely many choice of $\beta$, as long as it satisfies[3]

$$\frac{1-\beta_{k+1}}{\beta_{k+1}^2} \leq \frac{1}{\beta_k^2}.$$

- What we want: taking $\lim\limits_{\alpha\to 0}$ not to remove $\ddot{\mathbf{X}}$ or blow up the ODE. Try

$$\beta_k = \frac{k}{k+3}.$$

$$\beta_{k-1} = \frac{k-1}{k+2} = 1 - \frac{-3}{k+2} \overset{k\gg 2}{\approx} 1 - \frac{3}{k} \overset{k=\frac{t}{\sqrt{\alpha}}}{=} 1 - \frac{3\sqrt{\alpha}}{t}.$$

Now the mysterious 3 appears.

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

# Finishing the derivation

- For $\beta_{k-1} = 1 - \dfrac{3\sqrt{\alpha}}{t}$,

  $$\frac{\alpha}{2}(1+\beta_{k-1})\ddot{\mathbf{X}} + \sqrt{\alpha}(1-\beta_{k-1})\dot{\mathbf{X}} + \alpha\nabla f(\mathbf{X}) + o(\sqrt{\alpha}) = 0$$

  becomes

  $$\frac{\alpha}{2}\Big(2 - \frac{3\sqrt{\alpha}}{t}\Big)\ddot{\mathbf{X}} + \sqrt{\alpha}\Big(\frac{3\sqrt{\alpha}}{t}\Big)\dot{\mathbf{X}} + \alpha\nabla f(\mathbf{X}) + o(\sqrt{\alpha}) = 0.$$

- Divide the whole equation by $\alpha$, and note that $o(\sqrt{\alpha})$ contains terms with cubic power or higher in $\sqrt{\alpha}$ and hence they have $\alpha$,

  $$\frac{1}{2}\Big(2 - \frac{3\sqrt{\alpha}}{t}\Big)\ddot{\mathbf{X}} + \frac{3}{t}\dot{\mathbf{X}} + \nabla f(\mathbf{X}) + o(\sqrt{\alpha}) = 0.$$

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

# Why $h = \alpha$ does not work

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

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

- Take $\lim_{\alpha\to 0}$ makes the whole equation disappear.

# Last page - summary

- Nesterov's accelerated gradient (NAG)

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

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

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

- Standard ODE theory (not discussed here) gives

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

  As ODE $\iff$ NAG, this partially explains

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

End of document