Convergence Analysis of ODE Models for Accelerated First-Order Methods via Positive Semidefinite Kernels

Jungbin Kim **Insoon Yang**

Department of Electrical and Computer Engineering
Seoul National University

NeurIPS 2023

ODE Models in Convex Optimization

$$\min_{x \in \mathbb{R}^d} f(x)$$

Nesterov's accelerated gradient method (AGM):[1]

$$y_{k+1} = x_k - s\nabla f(x_k)$$
$$x_{k+1} = y_{k+1} + \frac{k-1}{k+2}(y_{k+1} - y_k).$$

Continuous-time limit of AGM:[2]

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

Goal: Develop a systematic methodology for analyzing convergence rates of ODE models.

[1] Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$".

[2] Su, Boyd, and Candes, "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights".

Discrete-Time PEP (Drori and Teboulle)

General form of first-order methods:

$$x_{k+1} = x_k - \sum_{j=0}^{k} h_{k,j} \nabla f(x_j), \tag{1}$$

parametrized by the coefficients $\{h_{k,j}\}$.

Proving convergence rate of (1)

$\Updownarrow$

Veryfing positive semidefiniteness of matrix[a]

[a] $x^\top M x \geq 0 \ \forall x$.

Drori and Teboulle, "Performance of first-order methods for smooth convex minimization: A novel approach".

Continuous-Time PEP (Ours)

General form of continuous-time models:

$$\dot{X}(t) = -\int_0^t H(t,\tau)\nabla f(X(\tau))\,d\tau, \tag{2}$$

parametrized by the **H-kernel** $H(t,\tau)$.[4]

Proving convergence rate of (2)

$\Updownarrow$

Veryfing positive semidefiniteness of integral kernel[a]

[a] $\iint k(t,\tau)f(t)f(\tau)\,dtd\tau \geq 0\ \forall f$.

[4] Kim and Yang, "Unifying Nesterov's Accelerated Gradient Methods for Convex and Strongly Convex Objective Functions".

Continuous PEP for Minimizing Function Values

$$\dot{X}(t) = -\int_0^t H(t,\tau)\nabla f(X(\tau))\, d\tau \tag{2}$$

Theorem (Function Value PEP)

Given $\nu > 0$, Lagrange multiplier $\lambda(t)$. Then, (2) achieves

$$f(X(T)) - f(x^*) \leq \nu \|x_0 - x^*\|^2,$$

if the following symmetric **PEP kernel** *is positive semidefinite:*

$$S(t,\tau) = \nu \left(\lambda(t) H(t,\tau) + \dot{\lambda}(t) \int_\tau^t H(s,\tau)\, ds \right) - \frac{1}{2} \dot{\lambda}(t) \dot{\lambda}(\tau),\ t \geq \tau.$$

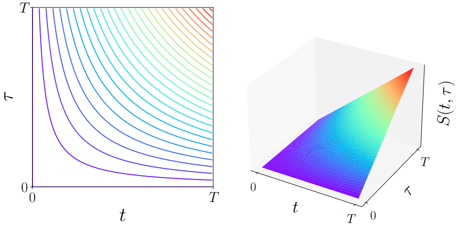
- Can be extended to strongly convex case ($\mu > 0$).

Continuous PEP for Minimizing Function Values

AGM ODE:

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0$$
$$\Leftrightarrow \dot{X}(t) = -\int_0^t \frac{\tau^3}{t^3} \nabla f(X(\tau))\, d\tau.$$

With $\lambda(t) = \frac{t^2}{T^2}$, we have $S(t,\tau) = \left(\nu - \frac{2}{T^2}\right)\frac{t\tau}{T^2} \succeq 0$ when $\nu \geq \frac{2}{T^2}$.



$$f(X(T)) - f(x^*) \leq \frac{2}{T^2}\|x_0 - x^*\|^2.$$

Continuous PEP for Minimizing Gradient Norms

$$\dot{X}(t) = -\int_0^t H(t,\tau)\nabla f(X(\tau))\,d\tau \tag{2}$$

Theorem (Gradient Norm PEP)

Given $\nu > 0$, Lagrange multiplier $\lambda(t)$. Then, (2) achieves

$$\|\nabla f(X(T))\|^2 \leq 4\nu(f(x_0) - f(x^*)),$$

if the following symmetric **PEP kernel** *is positive semidefinite:*

$$S(t,\tau) = \nu\left(\frac{H(t,\tau)}{\lambda(\tau)} + \frac{\dot{\lambda}(\tau)}{\lambda(\tau)^2}\int_\tau^t H(t,s)\,ds\right) - \frac{\dot{\lambda}(t)\dot{\lambda}(\tau)}{2\lambda(t)^2\lambda(\tau)^2},\ t \geq \tau.$$

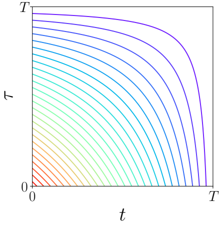
- Can be extended to strongly convex case ($\mu > 0$).
- Can also prove convergence rates on $\|\dot{X}(T)\|^2$.

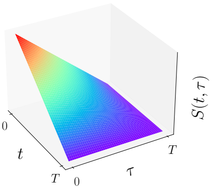
Continuous PEP for Minimizing Gradient Norms

OGM-G ODE:[5]

$$\ddot{X}(t) + \frac{3}{T-t}\dot{X}(t) + \nabla f(X(t)) = 0$$
$$\Leftrightarrow \dot{X}(t) = -\int_0^t \frac{(T-t)^3}{(T-\tau)^3}\nabla f(X(\tau))\,d\tau.$$

With $\lambda(t) = \frac{T^2}{(T-t)^2}$, we have $S(t,\tau) = \left(\nu - \frac{2}{T^2}\right)\frac{(T-t)(T-\tau)}{T^2} \succeq 0$ when $\nu \geq \frac{2}{T^2}$.





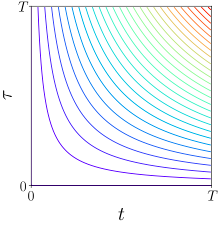
$$\|\nabla f(X(T))\|^2 \leq \frac{8}{T^2}(f(x_0) - f(x^*)).$$

[5] Suh, Roh, and Ryu, "Continuous-Time Analysis of Accelerated Gradient Methods via Conservation Laws in Dilated Coordinate Systems".

Correspondence Between Minimizing Function Values and Minimizing Gradient Norms

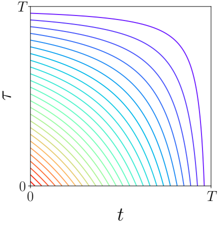
$$\dot{X}(t) = -\int_0^t \frac{\tau^3}{t^3} \nabla f(X(\tau))\, d\tau$$

$$S^F(t,\tau) = \left(\nu - \frac{2}{T^2}\right) \frac{t\tau}{T^2}$$



$$\dot{X}(t) = -\int_0^t \frac{(T-t)^3}{(T-\tau)^3} \nabla f(X(\tau))\, d\tau$$

$$S^G(t,\tau) = \left(\nu - \frac{2}{T^2}\right) \frac{(T-t)(T-\tau)}{T^2}$$



Anti-transpose relationships:

$$H^F(t,\tau) = H^G(T-\tau, T-t)$$
$$S^F(t,\tau) = S^G(T-\tau, T-t)$$

Correspondence Between Minimizing Function Values and Minimizing Gradient Norms

$$\dot{X}(t) = -\int_0^t H^F(t,\tau)\nabla f(X(\tau))\,d\tau \tag{F}$$

$$\dot{X}(t) = -\int_0^t H^G(t,\tau)\nabla f(X(\tau))\,d\tau \tag{G}$$

Theorem (Correspondence between F and G)

If $H^F(t,\tau) = H^G(T-\tau, T-t)$, then the following are equivalent:

- (F) *achieves* $f(X(T)) - f(x^*) \le \nu\|x_0 - x^*\|^2$.
- (G) *achieves* $\|\nabla f(X(T))\|^2 \le 4\nu(f(x_0) - f(x^*))$.

- Can be extended to strongly convex case ($\mu > 0$).

Conclusion

Contributions

We introduced **Continuous PEP**, a systematic methodology for analyzing ODE models in convex optimization.

- Enhances the understanding of continuous-time analysis.
- Unlocks new opportunities for studying discrete-time PEP.

Acknowledgement: This work was supported in part by the National Research Foundation of Korea funded by MSIT(2020R1C1C1009766), and the Information and Communications Technology Planning and Evaluation (IITP) grant funded by MSIT(2022-0-00124, 2022-0-00480).