Unifying Nesterov's Accelerated Gradient Methods for Convex and Strongly Convex Objective Functions

Jungbin Kim **Insoon Yang**

Department of Electrical and Computer Engineering
Seoul National University

ICML 2023

Nesterov Acceleration for Convex Optimization

Problem setting:

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

where $f:\mathbb{R}^n\to\mathbb{R}$ is an L-smooth and (μ-strongly) convex function.

Nesterov's accelerated gradient method (AGM):

AGM-C (convex)

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

AGM-SC (strongly convex)

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$$
$$y_{k+1} = x_{k+1} + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(x_{k+1}-x_k)$$

- AGM-SC with $\mu = 0$ is not equivalent to AGM-C.
- We present a unified framework for resolving the inconsistency.

Inconsistency Between Nesterov's AGM

ODE models: Discrete-time $\rightarrow$ Continuous-time.

- Su *et al.* (2016): $\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0$ (AGM-C ODE).
- Wilson *et al.* (2021): $\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0$ (AGM-SC ODE).

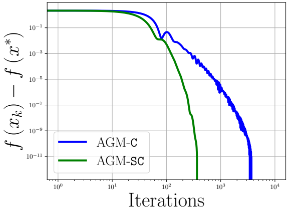
AGM-SC ODE with $\mu = 0$ is not equivalent to AGM-C ODE.

Lagrangian formulations: ODE models can be obtained from the Euler-Lagrange equation $\frac{d}{dt}\frac{\partial}{\partial \dot{X}}\mathcal{L}(X, \dot{X}, t) = \frac{\partial}{\partial X}\mathcal{L}(X, \dot{X}, t)$.

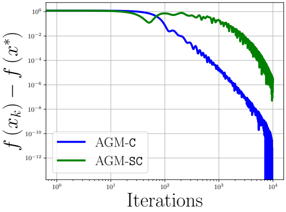
- Wibisono *et al.* (2016): First Bregman Lagrangian (convex).
- Wilson *et al.* (2021): Second Bregman Lagrangian (strongly convex).

Second Lagrangian with $\mu = 0$ is not equivalent to First Lagrangian.

Inconsistency Between Nesterov's AGM



Large μ



Small μ

- AGM-SC $>$ AGM-C for large μ.
- AGM-C $>$ AGM-SC for small μ.

Q) **Is there a unified algorithm that combines AGM-C and AGM-SC?**

Our Contributions

Q) Is there a unified algorithm that combines AGM-C and AGM-SC?

We propose:

- **Unified Bregman Lagrangian** to combine two known Lagrangians.
- **Unified AGM ODE** to combine two known ODE models.
- **Unified AGM** to combine two algorithms: AGM-C and AGM-SC.
- **Unified ATM**, a unified accelerated higher-order gradient method.
- **Unified AGM-G ODE**, a novel ODE model for minimizing gradient norm of strongly convex objective functions.

Main Result 1. Unified Lagrangian Formulation

Given smooth real functions $\alpha, \beta, \gamma : \mathbb{R} \to \mathbb{R}$,

Unified Bregman Lagrangian:

$$\mathcal{L}\left(X, \dot{X}, t\right) = e^{\alpha+\gamma}\left(\left(1 + \mu e^{\beta}\right) D_h\left(X + e^{-\alpha}\dot{X}, X\right) - e^{\beta} f(X)\right).$$

Unified Bregman Lagrangian flow (from Euler–Lagrange equation):

$$\begin{aligned} \dot{X} &= e^{\alpha}(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= \frac{\mu\dot{\beta}e^{\beta}}{1 + \mu e^{\beta}}\left(\nabla h(X) - \nabla h(Z)\right) - \frac{e^{\alpha+\beta}}{1 + \mu e^{\beta}}\nabla f(X). \end{aligned}$$

Theorem (Convergence of Unified Bregman Lagrangian flow)

$$f(X(t)) - f\left(x^*\right) \leq O\left(e^{-\beta(t)}\right).$$

Main Result 1. Unified Lagrangian Formulation

First Bregman Lagrangian flow (convex case):

$$\frac{d}{dt}\nabla h(Z) = -e^{\alpha+\beta}\nabla f(X).$$

Second Bregman Lagrangian flow (strongly convex case):

$$\frac{d}{dt}\nabla h(Z) = \dot{\beta}\left(\nabla h(X) - \nabla h(Z)\right) - \frac{e^{\alpha}}{\mu}\nabla f(X).$$

Unified Bregman Lagrangian flow:

$$\frac{d}{dt}\nabla h(Z) = \frac{\mu\dot{\beta}e^{\beta}}{1+\mu e^{\beta}}\left(\nabla h(X) - \nabla h(Z)\right) - \frac{e^{\alpha+\beta}}{1+\mu e^{\beta}}\nabla f(X).$$

Unified Bregman Lagrangian flow reduces to:

- First Bregman Lagrangian flow (Wibisono *et al.*, 2016) when $\mu = 0$.
- Second Bregman Lagrangian flow (Wilson *et al.*, 2021) as $t \to \infty$.

Main Result 2. Unified ODE Model

Choosing $\beta(t) = \log\left(\frac{1}{\mu}\sinh^2\left(\frac{\sqrt{\mu}}{2}t\right)\right)$, $\alpha(t) = \dot{\beta}(t)$, and $\gamma(t) = \beta(t)$,

Unified AGM ODE:

$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}t\right) + \frac{3\sqrt{\mu}}{2}\coth\left(\frac{\sqrt{\mu}}{2}t\right)\right)\dot{X} + \nabla f(X) = 0$$

Theorem (Convergence of Unified AGM ODE)

$$f(X(t)) - f(x^*) \leq O\left(\min\left\{1/t^2, e^{-\sqrt{\mu}t}\right\}\right).$$

Main Result 2. Unified ODE Model

AGM-C ODE, $O(1/t^2)$ rate:

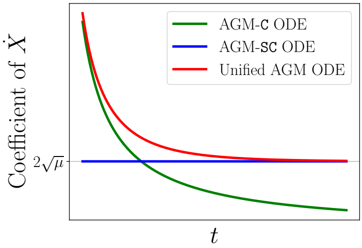
$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0.$$

AGM-SC ODE, $O(e^{-\sqrt{\mu}t})$ rate:

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0.$$

Unified AGM ODE, $O(\min\{1/t^2, e^{-\sqrt{\mu}t}\})$ rate:

$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}t\right) + \frac{3\sqrt{\mu}}{2}\coth\left(\frac{\sqrt{\mu}}{2}t\right)\right)\dot{X} + \nabla f(X) = 0$$



Main Result 3. Unified Algorithm

Unified AGM:

$$x_{k+1} = y_k - \frac{1}{L}\nabla f\left(y_k\right)$$

$$y_{k+1} = x_{k+1} + \frac{\left(\tanh\left(\frac{k+1}{2}\iota\sqrt{q}\right) - \sqrt{q}\right)\left(\coth\left(\frac{k+2}{2}\iota\sqrt{q}\right) - \sqrt{q}\right)}{1-q}\left(x_{k+1} - x_k\right),$$

where $q = \mu/L$ and $\iota = -\frac{\log\left(1-\sqrt{q}\right)}{\sqrt{q}}$.

Theorem (Convergence of Unified AGM)

$$f(x_k) - f\left(x^*\right) \leq O\left(\min\left\{1/k^2, \left(1-\sqrt{\mu/L}\right)^k\right\}\right).$$

Main Result 3. Unified Algorithm

AGM-C (convex)

$$x_{k+1} = y_k - \frac{1}{L}\nabla f\left(y_k\right)$$
$$y_{k+1} = x_{k+1} + \frac{k-1}{k+2}\left(x_{k+1} - x_k\right)$$

AGM-SC (strongly convex)

$$x_{k+1} = y_k - \frac{1}{L}\nabla f\left(y_k\right)$$
$$y_{k+1} = x_{k+1} + \frac{1-\sqrt{q}}{1+\sqrt{q}}\left(x_{k+1} - x_k\right)$$

Unified AGM

$$x_{k+1} = y_k - \frac{1}{L}\nabla f\left(y_k\right)$$
$$y_{k+1} = x_{k+1} + \frac{\left(\tanh\left(\frac{k+1}{2}\iota\sqrt{q}\right) - \sqrt{q}\right)\left(\coth\left(\frac{k+2}{2}\iota\sqrt{q}\right) - \sqrt{q}\right)}{1-q}\left(x_{k+1} - x_k\right),$$

Unified AGM reduces to:

- AGM-C and $O(1/k^2)$ rate when $\mu = 0$.
- AGM-SC and $O((1-\sqrt{\mu/L})^k)$ rate as $k \to \infty$.

Main Result 4. Extension to Higher-Order Setting

Problem setting:

- Distance-generating funtion h satisfies $D_h(x,y) \geq \frac{1}{p}\|y-x\|^p$.
- f is L-smooth of order $p-1$: $\|\nabla^{p-1}f(y) - \nabla^{p-1}f(x)\| \leq L\|y-x\|$.
- f is μ-uniformly convex with respect to h: $\mu D_h(x,y) \leq D_f(x,y)$.

We propose **Unified accelerated tensor method (Unified ATM)**.

Theorem (Convergence of Unified ATM)

$$f(X(t)) - f(x^*) \leq O\left(\min\left\{1/t^p, e^{-p\sqrt[p]{C\mu}t}\right\}\right),$$

$$f(x_k) - f(x^*) \leq O\left(\min\left\{1/k^p, \left(1+p\sqrt[p]{C\mu/L}\right)^{-k}\right\}\right).$$

This extends the $O(1/t^p)$ and $O(1/k^p)$ convergence rate results for the convex case ($\mu = 0$), established in (Wibisono *et al., 2016).*

Main Result 5. Gradient Norm Minimization

H-kernel (novel tool): $\dot{X}(t) = -\int_0^t H(t,\tau)\nabla f(X(\tau))\,d\tau$.

AGM-C ODE (Su *et al.*, 2016), $f(X(T)) - f(x^*) \le O(1/T^2)$:

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

OGM-G ODE (Suh *et al.*, 2022), $\|\nabla f(X(T))\|^2 \le O(1/T^2)$:

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

Symmetric relationships:

- Time-reversed relationship ($t \leftrightarrow T-t$) between the coefficients of $\dot{X}$.
- Anti-transpose relationship ($t \leftrightarrow T-\tau$) between the "H-kernel"s.

Main Result 5. Gradient Norm Minimization

Symmetric relationships:

- Time-reversed relationship ($t \leftrightarrow T - t$) between the coefficients of $\dot{X}$.
- Anti-transpose relationship ($t \leftrightarrow T - \tau$) between the "$H$-kernel"s.

Unified AGM ODE, $f(X(T)) - f(x^*) \le O(\min\{1/T^2, e^{-\sqrt{\mu}T}\})$:

$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2} \tanh\left(\frac{\sqrt{\mu}}{2} t \right) + \frac{3\sqrt{\mu}}{2} \coth\left(\frac{\sqrt{\mu}}{2} t \right) \right) \dot{X} + \nabla f(X) = 0$$

Unified AGM-G ODE (from symmetric relationships):

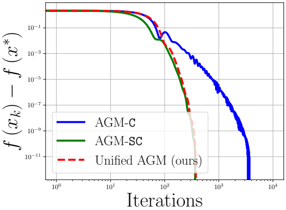
$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2} \tanh\left(\frac{\sqrt{\mu}}{2} (T-t) \right) + \frac{3\sqrt{\mu}}{2} \coth\left(\frac{\sqrt{\mu}}{2} (T-t) \right) \right) \dot{X} + \nabla f(X) = 0$$

Theorem (Convergence of Unified AGM-G ODE)

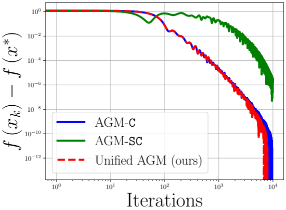
$$\|\nabla f(X(T))\|^2 \le O\left(\min\left\{ 1/T^2, e^{-\sqrt{\mu}T} \right\} \right).$$

Numerical Experiment: ℓ_2-Regularized Logistic Regression

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m}\left(\sum_{i=1}^{m}(-y_i a_i^T x + \log(1 + e^{a_i^T x})) + \lambda \|x\|^2\right)$$



Large μ



Small μ

- Unified AGM $\approx$ AGM-SC $>$ AGM-C for large μ.
- Unified AGM $\approx$ AGM-C $>$ AGM-SC for small μ.

Unified AGM combines the benefits of AGM-C and AGM-SC.

Conclusion

Contributions

We developed a framework for designing algorithms that handle the convex case ($\mu = 0$) and the strongly convex case ($\mu > 0$) in a unified way.

- Unified Bregman Lagrangian, Unified AGM ODE, Unified AGM.
- Extension to higher-order setting: Unified ATM.
- Gradient norm minimization: Unified AGM-G.

Acknowledgement: This work was supported in part by Samsung Electronics, 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).