Time-Reversed Dissipation Induces Duality Between Minimizing Gradient Norm and Function Value

Jaeyeon Kim[†], Asuman Ozdaglar[‡], Chanwoo Park[‡], Ernest K.Ryu[†]

[†]: SNU Mathematics , [‡]: MIT EECS

2023. 07. 29

Duality between methods

$f(x)$ reducing method $\xleftrightarrow{\textit{Duality}}$ $\|\nabla f(x)\|^2$ reducing method

- We observed a **symmetric phenomenon** between pairs of **first-order methods.**
  - They have similar coefficients.
  - They share the convergence rate.
- Can we find a **duality framework** that generalizes this phenomenon?

First Order Methods and H-dual

Definition

f is L-smooth convex. An N-step First Order Methods with $H = \{h_{k,i}\}_{0 \le i < k \le N}$ is:

$$x_{k+1} = x_k - \frac{1}{L} \sum_{k=0}^{i} h_{k+1,i} \nabla f(x_i), \quad k = 0, 1, \ldots, N-1.$$

Its **H-dual** is defined as

$$x_{k+1} = x_k - \frac{1}{L} \sum_{k=0}^{i} h_{N-i,N-1-k} \nabla f(x_i), \quad k = 0, 1, \ldots, N-1,$$

$$\begin{bmatrix} h_{10} & 0 & \cdots & 0 \\ h_{20} & h_{21} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n0} & h_{n1} & \cdots & h_{n(n-1)} \end{bmatrix} \xrightarrow{\text{Anti-Transpose}} \begin{bmatrix} h_{n(n-1)} & 0 & \cdots & 0 \\ h_{n(n-2)} & h_{(n-1)(n-2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n0} & h_{(n-1)0} & \cdots & h_{10} \end{bmatrix}$$

Example 1 : OGM and OGM-G

(OGM)[1] and $(OGM\text{-}G)$[2]:[3]

$$x_{k+1} = x_k - \frac{1}{L}\sum_{i=0}^{k}\left(\frac{\theta_i^2(2\theta_i - 1)}{\theta_k^2\theta_{k+1}} + \delta_{k,i}\right)\nabla f(x_i), \qquad (OGM)$$

$$y_{k+1} = y_k - \frac{1}{L}\sum_{i=0}^{k}\left(\frac{\theta_{N-k-1}^2(2\theta_{N-k-1} - 1)}{\theta_{N-i-1}^2\theta_{N-i}} + \delta_{N-k-1,N-i-1}\right)\nabla f(x_i) \qquad (OGM\text{-}G)$$

- They are **H-dual** of each other.
- Convergence rates:

$$f(x_N) - f_\star \le \frac{1}{\theta_N^2}\frac{L}{2}\|x_0 - x_\star\|^2 \quad (OGM),$$

$$\frac{1}{2L}\|\nabla f(y_N)\|^2 \le \frac{1}{\theta_N^2}(f(y_0) - f_\star) \quad (OGM\text{-}G).$$

[1] [Kim and Fessler, 2016]
[2] [Kim and Fessler, 2021]
[3] $\theta_{-1} = \theta_0 = 1$, $\theta_{i+1}^2 - \theta_{i+1} = \theta_i^2$ for $0 \le i \le N-2$, $\theta_N^2 - \theta_N = 2\theta_{N-1}^2$.

Aim

These three pairs of methods share the same factor of convergence rate and their H-matrices are in the Anti-Transpose relationship.

$$\Big[(OGM),(OGM\text{-}G)\Big],\quad \Big[(GD),(GD)\Big],\quad \Big[(OBL\text{-}F_\flat),(OBL\text{-}G_\flat)\Big]$$ [4]

Main theorem:

A method reduces *function value* $\leftrightarrow$ Its **H-dual** reduces *gradient norm*

[4] [Park and Ryu, 2021]

Energy function Structure and H-duality

$$
\begin{aligned}
x_\star :&= \mathrm{argmin}\, f(x). \\
f_\star :&= f(x_\star). \\
[\![x, y]\!] :&= f(y) - f(x) + \langle \nabla f(y), x - y\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \le 0. \\
\mathcal{U}_k :&= \frac{L}{2}\|x_0 - x_\star\|^2 + \sum_{i=0}^{k-1} u_i [\![x_i, x_{i+1}]\!] + \sum_{i=0}^{k} (u_i - u_{i-1})[\![x_\star, x_i]\!], \\
\mathcal{V}_k :&= v_0\Big([\![y_N, y_\star]\!] + f(y_0) - f_\star\Big) + \sum_{i=0}^{k-1} v_{i+1}[\![y_i, y_{i+1}]\!] + \sum_{i=0}^{k-1} (v_{i+1} - v_i)[\![y_N, y_i]\!].
\end{aligned}
$$

$u_{-1} = 0$. $\{u_i\}_{i=0}^N$, $\{v_i\}_{i=0}^N$ are free variables. If they're positive and *monotonically increasing*, $\{\mathcal{U}_k\}$ and $\{\mathcal{V}_k\}$ are *monotonically decreasing*.

Energy function structure and H-duality

If

$$u_N(f(x_N) - f_\star) \le \mathcal{U}_N, \tag{C1}$$

$$f(x_N) - f_\star \le \frac{\mathcal{U}_N}{u_N} \le \cdots \le \frac{\mathcal{U}_{-1}}{u_N} = \frac{1}{u_N}\frac{L}{2}\|x_0 - x_\star\|^2.$$

If

$$\frac{1}{2L}\|\nabla f(y_N)\|^2 \le \mathcal{V}_N, \tag{C2}$$

$$\frac{1}{2L}\|\nabla f(y_N)\|^2 \le \mathcal{V}_N \le \cdots \le \mathcal{V}_0 \le v_0\left(f(y_0) - f_\star\right).$$

Theorem

Assume $v_i = \frac{1}{u_{N-i}} > 0$ *for* $i = 0, \ldots, N$.

$$[\text{(C1)}] \quad \Leftrightarrow \quad [\text{(C2) } \textit{for the } \textbf{\textit{H-dual}}].$$

Time-Reversed Dissipation

- Continuous-time limit $N \to \infty$ of first-order method becomes ODE.
- The notion of **H-dual** becomes:

$$\ddot{X}(t) + \gamma'(t)\dot{X}(t) + \nabla f(X(t)) = 0, \quad \ddot{Y}(t) + \gamma'(T-t)\dot{Y}(t) + \nabla f(Y(t)) = 0.$$

- Friction terms are time-reversed.

Applications of H-duality

- Constructing *gradient-norm* reducing method is quite hard rather than constructing the *function-value* reducing method.
- Utilizing **H-duality** makes it easier.
- In the composite minimization setting, we obtain *(Super FISTA-G)* (*SFG*), which reduces $\min_{v \in \partial F(x)} \|v\|^2$.
  - $F := f + h$, h is convex.
  - $h = 0$: $\min_{v \in \partial F(x)} \|v\|^2 = \|\nabla f(x)\|^2$.
- (*SFG*) is state-of-the-art : 5.26 times faster than *(FISTA-G)*[5].

[5] [Lee et al., 2021]

Conclusion

- In this work, we defined the notion of **H-duality**.
- We established that the **H-dual** of a method reducing *function value* is another method reducing *gradient norm*.
- We are currently working on further generalizing **H-duality** in the *Mirror Descent* setting.

References

Drori, Y. (2017).
The exact information-based complexity of smooth convex minimization.
Journal of Complexity, 39:1–16.

Drori, Y. and Teboulle, M. (2014).
Performance of first-order methods for smooth convex minimization: a novel approach.
Mathematical Programming, 145(1):451–482.

Kim, D. and Fessler, J. A. (2016).
Optimized first-order methods for smooth convex minimization.
Mathematical Programming, 159(1):81–107.

Kim, D. and Fessler, J. A. (2021).
Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions.
Journal of Optimization Theory and Applications, 188(1):192–219.

Lee, J., Park, C., and Ryu, E. K. (2021).
A geometric structure of acceleration and its role in making gradients small fast.
NeurIPS.

Park, C. and Ryu, E. K. (2021).
Optimal first-order algorithms as a function of inequalities.
arXiv preprint arXiv:2110.11035.

Concluding Remark : (*OGM*) and (*OGM-G*)

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}}(x_k^+ - x_{k-1}^+) + \frac{\theta_k}{\theta_{k+1}}(x_k^+ - x_k) \qquad (OGM)$$

$$y_{k+1} = y_k^+ + \frac{(\theta_{N-k} - 1)(2\theta_{N-k-1} - 1)}{2\theta_{N-k}(2\theta_{N-k} - 1)}(y_k^+ - y_{k-1}^+) + \frac{2\theta_{N-k-1} - 1}{2\theta_{N-k} - 1}(y_k^+ - y_k) \qquad (OGM\text{-}G)$$

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}}\left(x_K^+ - x_{k-1}^+\right) \qquad (Nesterov\text{'s } FGM)$$

(*OGM*) is exactly optimal [6].

[6] [Drori, 2017]

Concluding Remark : (*SFG*)

$$y_{k+1} = y_k^{\oplus,4} + \frac{(N-k+1)(2N-2k-1)}{(N-k+3)(2N-2k+1)}\left(y_k^{\oplus,4} - y_{k-1}^{\oplus,4}\right) + \frac{(4N-4k-1)(2N-2k-1)}{6(N-k+3)(2N-2k+1)}\left(y_k^{\oplus,4} - y_k\right)$$

$$y_N = y_{N-1}^{\oplus,4} + \frac{3}{10}\left(y_{N-1}^{\oplus,4} - y_{N-2}^{\oplus,4}\right) + \frac{3}{40}\left(y_{N-1}^{\oplus,4} - y_{N-1}\right) \qquad (SFG)$$

for $k = 0, \ldots, N-2$, where

Theorem

(*SFG*) *exhibits*

$$\min_{v \in \partial F(y_N^{\oplus,4})} \|v\|^2 \leq \frac{50L}{(N+2)(N+3)}\left(F(y_0) - F_\star\right).$$

(*SFG*) is the state-of-the-art : 5 times faster than *FISTA-G* [7].

[7] [Lee et al., 2021]

Example 2 : Gradient Descent

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \quad k = 0, \dots, N-1. \tag{GD}$$

- (GD) is **H-dual** of itself.
- Convergence rate[8]:

$$f(x_N) - f_\star \le \frac{1}{2N+1}\frac{L}{2}\|x_0 - x_\star\|^2,$$
$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \le \frac{1}{2N+1}(f(x_0) - f_\star).$$

[8] [Drori and Teboulle, 2014, Kim and Fessler, 2021]