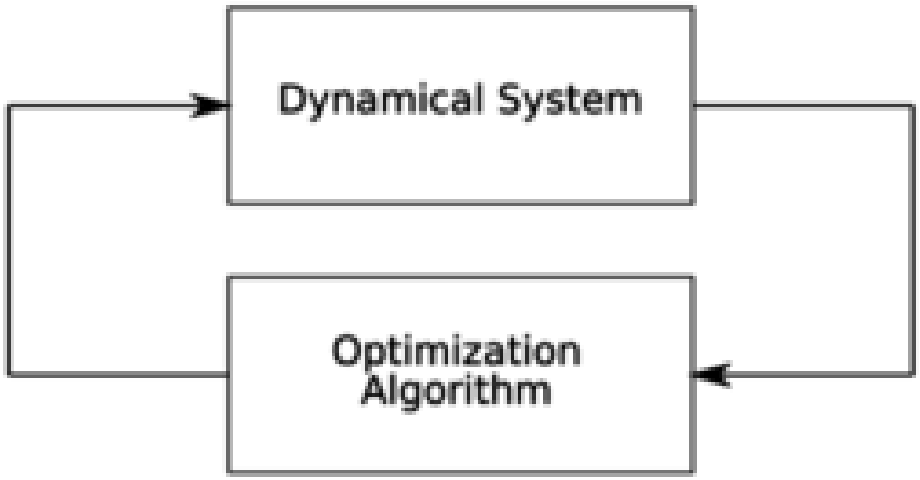
Analysis of Some First Order Optimization Algorithms by Dissipativity Theory

Majid Darehmiraki

24 May 2022

Behbahan Khatam Alanbia university of Technology, Khouzestan, Iran

In recent years, renewed attention has been paid to the fact that many numerical optimization algorithms can be interpreted as dynamical systems. This perspective is essential to bridge the gap between algorithms and their implementation as feedback systems.



Hauswirth, A., Bolognani, S., Hug, G., & Dorfler, F. (2021). Optimization algorithms as robust feedback controllers. arXiv preprint arXiv:2103.11329.

How to prove the stability of dynamical systems?

- Lyapunov function: A universal approach to analyzing the stability of dynamical systems is to construct a Lyapunov function that decreases along the trajectories of the system, proving asymptotic convergence.
- Integral quadratic constraints: A powerful LMI-based tool for stability analysis. The feasibility of LMI implies the linear convergence of the algorithm.
- Dissipativity theory: Dissipativity has been introduced by Willems and is motivated by the concept of passivity, a concept from electrical network theory which relates the stored energy in an electrical network with the supplied energy into the network.

In systems and control theory one often encounters nonlinear control systems described, in the state space form, by means of a set of ordinary differential equations of the following type:

$$\begin{aligned} \dot{x} &= f(x) + G(x)u, \\ y &= h(x). \end{aligned} \tag{1}$$

Definition

The control system (24) is said to be dissipative with respect to the supply rate $S : \mathcal{R}^n \times \mathcal{R}^p \times \mathcal{R}^q \to \mathcal{R}$, if there exists a positive semidefinite storage function $V : \mathcal{R}^n \to \mathcal{R}$ such that the (integral) dissipation inequality

$$V(x(t_1)) - V(x(t_2)) \leq \int_{t_0}^{t_1} S(x(t), u(t), y(t))dt \tag{2}$$

If the storage function V is smooth then the integral dissipation inequality (25) can be rewritten as

$$\dot{V}(x(t)) \leq S(x(t), u(t), y(t)). \tag{3}$$

Definition

The control system (24) with $p = q$ is said to be passive, if there exist a positive semidefinite storage function $V : \mathcal{R}^n \to \mathcal{R}$ such that the following dissipation inequality is satisfied:

$$\nabla f(x)(f(x) + G(x)u) \leq u^T h(x), \qquad \forall x,\ u. \tag{4}$$

Because (4) must hold for all u's, one obtains the so-called nonlinear positive real lemma

$$\nabla f(x) f(x) \leq 0$$
$$\nabla f(x) G(x) = h^T(x).$$

Linear Systems

The most well-studied class of control systems are linear time-invariant control systems given by

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx. \end{aligned} \tag{5}$$

For this class of systems, dissipativity theory can fully deploy its power because supply rates and storage functions can be computed efficiently.
A semidefinite program, which can be seen as a generalization of a linear program, is a convex optimization problem and has the form

$$\begin{aligned} &\textit{minimize} \quad c^T \xi \\ &F_0 + \sum_{i=1}^{k} \xi_i F_i \leq 0, \\ &D\xi = e \end{aligned} \tag{6}$$

$$v(x) = x^T P x \tag{7}$$

with $P \geq 0$. Then, for example, the dissipation inequality (4) for passivity for system (5) turns into

$$2x^T P(Ax + Bu) \leq x^T C^T u \tag{8}$$

or equivalently into

$$\begin{pmatrix} x \\ u \end{pmatrix}^T \begin{pmatrix} PA + A^T P & PB - \frac{1}{2}C^T \\ B^T P - \frac{1}{2}C & 0 \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \leq 0 \tag{9}$$

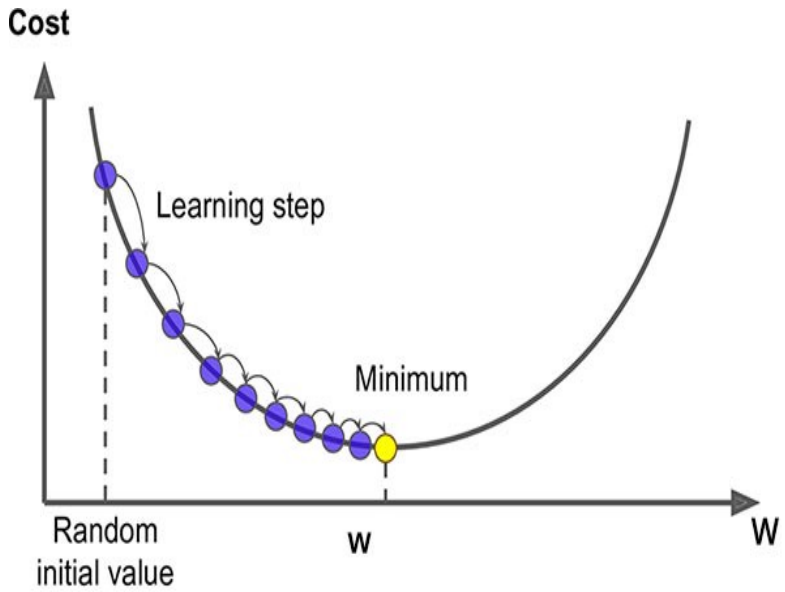
Optimization problems lie at the heart of many machine-learning formulations.

$$\begin{aligned} &\textit{minimize} && f(x) \\ &\textit{subject to} && x \in \Omega \end{aligned}$$

The simplest and probably most natural method for minimizing differentiable functions is gradient descent.

Gradient descent

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$



for constrained optimization, use projected gradient descent

$$x_{k+1} = \Pi_{\Omega}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

In real, it is the result of applying Euler's rule, with step-size $\alpha_k > 0$, to the gradient system

$$\dot{x} = -\nabla f(x)$$

to prevent oscillation, add a second order term

$$\ddot{x} = -b\dot{x} - \nabla f(x)$$
$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

heavy ball method with constants α, β

Canonical first order methods

Gradient

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

Heavy Ball

$$\begin{aligned} x_{k+1} &= y_k - \alpha \nabla f(x_k) \\ y_k &= (1+\beta)x_k - \beta x_{k-1} \end{aligned}$$

Nesterov

$$\begin{aligned} x_{k+1} &= y_k - \alpha \nabla f(y_k) \\ y_k &= (1+\beta)x_k - \beta x_{k-1} \end{aligned}$$

The main Dynamical system

$$x_{k+1} = Ax_k + Bu_k$$

The general form

$$y_{k+1} = y_k + \beta(y_k - y_{k-1}) - \alpha\nabla f(w_k)$$
$$w_k = y_k + \gamma(y_k - y_{k-1})$$

if set $x_k = [y_{k-1}^T \; y_k^T]^T$

$$A = \begin{pmatrix} 0 & I_d \\ -\beta I_d & (\beta+1)I_d \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ -\alpha I_d \end{pmatrix}$$

Classical Dissipativity Theory

Consider a linear dynamical system

$$x_{k+1} = Ax_k + Bu_k \tag{10}$$

Definition

The supply rate is a function $S : \mathcal{R}^n \times \mathcal{R}^m \to \mathcal{R}$ that maps any state/input pair (x, u) to a scalar measuring the amount of energy delivered from u to state x.

Definition

The dynamical system (10) is dissipative with respect to the supply rate S if there exists a function $V : \mathcal{R}^n \to \mathcal{R}^+$ such that $0 \leq V(x)$ for all $x \in \mathcal{R}^n$ and

$$V(x_{k+1}) - V(x_k) \leq S(x_k, u_k), \tag{11}$$

for all k. The function V is called a storage function, which quantifies the energy stored in the state x. In addition, (11) is called the dissipation inequality.

A variant of (11) known as the exponential dissipation inequality states that for some $0 \leq \rho < 1$, we have

$$V(x_{k+1}) - \rho^2 V(x_k) \leq S(x_k, u_k), \tag{12}$$

which states that at least a fraction $(1 - \rho^2)$ of the internal energy will dissipate at every step.

In summary, dissipativity involves three components

1. Positive semidefinite storage function V
2. the supply rate S
3. dissipation inequality

Theorem

Consider the following quadratic supply rate with $X \in \mathcal{R}^{(n+m)\times(n+m)}$ *and* $X = X^T$.

$$S(x,u) = \begin{pmatrix} x \\ u \end{pmatrix}^T X \begin{pmatrix} x \\ u \end{pmatrix}. \tag{13}$$

If there exists a matrix $P \in \mathcal{R}^{n+m}$ *with* $p \geq 0$ *such that*

$$\begin{pmatrix} A^TPA - \rho^2 P & A^TPB \\ B^TPA & B^TPB \end{pmatrix} - X \leq 0 \tag{14}$$

then the dissipation inequality (12) *holds for all trajectories of* (10) *with* $V(x) = x^TPx$.

For using the dissipativity theory for linear convergence rate analysis, we need basically two steps:

1. Choose a proper quadratic supply rate function S satisfying certain desired properties, e.g. $S(x_k, u_k) \leq 0$.
2. Solve the linear matrix inequality (14) to obtain a storage function V, which is then used to construct a Lyapunov function.

Dissipativity for Gradient Descent

Assume f is $L-$smooth, $m-$strongly convex and $\nabla f(y_*) = 0$.

$$y_{k+1} = y_k - \alpha \nabla f(y_k) \Longrightarrow y_{k+1} - y_* = y_k - y_* - \alpha f(y_k),$$

Define $x_k = y_k - y_*$ and $u_k = \nabla f(y_k)$ and set $A = I_p$ and $B = -\alpha I_p$.
In this case, the gradient descent method is modeled as (10).

Consider the following quadratic supply rate

$$S(x_k, u_k) = \begin{pmatrix} x_k \\ u_k \end{pmatrix}^T \begin{pmatrix} 2mLI_p & -(m+L)I_p \\ -(m+L)I_p & 2I_p \end{pmatrix} \begin{pmatrix} x_k \\ u_k \end{pmatrix} \qquad (15)$$

Due to co-coercivity $S(x_k, u_k) \leq 0$ for all k.
Set $P = p \otimes I_p$ and

$$V(x_k) = p\|x_k\|^2 = p\|y_k - y_*\|$$

The following LMI is obtained

$$\left(\begin{pmatrix} (1-\rho^2)p & -\alpha p \\ -\alpha p & \alpha^2\rho^2 \end{pmatrix} + \begin{pmatrix} -2mL & m+L \\ m+L & -2 \end{pmatrix} \right) \otimes I_p \leq 0$$

Therefor

$$p\|y_{k+1} - y_*\|^2 \leq \rho^2 p\|y_k - y_*\|^2$$

if there exist $p \geq 0$ such that

$$\begin{pmatrix} (1-\rho^2)p & -\alpha p \\ -\alpha p & \alpha^2 p \end{pmatrix} + \begin{pmatrix} -2mL & m+L \\ m+L & -2 \end{pmatrix} \leq 0 \qquad (16)$$

If we set $(\alpha, \rho, p) = (\frac{1}{L}, 1 - \frac{m}{L}, L^2)$, can recover the standard rate resulti in Ployak(1987).

Theorem

Let $x_* \in \operatorname{argmin}_{x \in \mathcal{R}^d} f(x)$ be a minimizer of $f : \mathcal{R}^d \to \mathcal{R} \cup \{\infty\}$ with a finite optimal value $f(x_*)$. Consider an iterative first order algorithm in the state-space form.

1. *Suppose the fixed points (ξ_*, u_*, y_*, x_*) of the state-space form satisfy*

$$\xi_* = A_k \xi_* + B_k u_*, \; y_* = C_k \xi_*, \; u_* = \phi(y_*), \; x_* = E_k \xi_* = y_* \text{ for all } k.$$

2. *Suppose there exist symmetric matrices M_k^1, M_k^2, M_k^3 such that the following inequalities hold for all k.*

$$\begin{aligned} F(x_{k+1}) - F(x_k) &\leq e_k^T M_k^1 e_k, \\ F(x_{k+1}) - F(x_*) &\leq e_k^T M_k^2 e_k, \\ 0 &\leq e_k^T M_k^3 e_k, \end{aligned}$$

   where $e_k = [(\xi_k - \xi_*)^T \; (u_k - u_*)^T]^T \in \mathcal{R}^{n+d}$ and M_k^3 is either zero or indefinite.

3. *Suppose there exists a nonnegative and nondecreasing sequence of real $\{a_k\}$, a sequence of nonnegative reals $\{\sigma_k\}$, and e sequence of $n \times n$ positive semidefinite matrices P_k satisfying*

$$M_k^0 + a_k M_k^1 + (a_{k+1} - a_k) M_k^2 + \sigma M_k^3 \leq 0 \text{ for all } k,$$

   where

$$M_k^0 = \begin{pmatrix} A_k^T P_{k+1} A_k - P_k & A_k^T P_{k+1} B_k \\ B_k^T P_{k+1} B_k & A_k^T P_{k+1} B_k \end{pmatrix}.$$

Then the sequence $\{x_k\}$ satisfies

Dissipativity for Nesterov's Method

Consider following dynamical system

$$\begin{aligned} y_{k+1} &= w_k - \alpha \nabla f(w_k) \\ w_k &= (1+\beta) y_k - \beta y_{k-1} \end{aligned} \tag{17}$$

Equations (17) can be rewritten as follows:

$$\begin{pmatrix} y_{k+1} - y_* \\ y_k - y_* \end{pmatrix} = A \begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \end{pmatrix} + B u_k \tag{18}$$

where $u_k = \nabla f(w_k)$, $A = \bar{A} \otimes I_p$, $B = \bar{B} \otimes I_p$ and

$$\bar{A} = \begin{pmatrix} 1+\beta & -\beta \\ 1 & 0 \end{pmatrix}, \qquad \bar{B} = \begin{pmatrix} -\alpha \\ 0 \end{pmatrix}.$$

If set

$$x_k = \begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \end{pmatrix}$$

Nesterov's accelerated method can be written in the form of (10).

Lemma

Let f be $L-$smooth and $m-$strongly convex with $m > 0$. Let y_* be the unique point satisfying $\nabla f(y_*) = 0$. Consider Nesterov's method (17) or equivalently (18). The following inequalities hold for all trajectories.

$$\begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \\ \nabla f(w_k) \end{pmatrix}^T X_1 \begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \\ \nabla f(w_k) \end{pmatrix} \leq f(y_k) - f(y_{k+1}),$$

$$\begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \\ \nabla f(w_k) \end{pmatrix}^T X_2 \begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \\ \nabla f(w_k) \end{pmatrix} \leq f(y_*) - f(y_{k+1})$$

where $X_i = \bar{X}_i \otimes I_p$ for $i = 1, 2$, and $\bar{X}_i$ are defined by

$$\bar{X}_1 = \frac{1}{2} \begin{pmatrix} \beta^2 m & -\beta^2 m & -\beta \\ -\beta^2 m & \beta^2 m & \beta \\ -\beta & \beta & \alpha(2 - L\alpha) \end{pmatrix} \tag{19}$$

$$\bar{X}_2 = \frac{1}{2} \begin{pmatrix} (1+\beta^2)m & -\beta(1+\beta)m & -(1+\beta) \\ -\beta(1+\beta)m & \beta^2 m & \beta \\ -(1+\beta) & \beta & \alpha(2 - L\alpha) \end{pmatrix} \tag{20}$$

One can define the supply rate as (13) with a particular choice of $X = \rho^2 X_1 + (1 - \rho^2)X_2$, $(0 < \rho < 1)$. Then this supply rate satisfies the condition

$$S(x_k, u_k) \leq \rho^2 (f(y_k) - f(y_*) - (f(y_{k+1} - f(y_*)). \tag{21}$$

Theorem

Let f be $L-$smooth and $m-$strongly convex with $m > 0$. Let y_* be the unique point satisfying $\nabla f(y_*) = 0$. Consider Nesterov's accelerated method (17). Set $\bar{X} = \rho^2 \bar{X}_1 + (1-\rho^2)\bar{X}_2$. Thus there exist a matrix $0 < \bar{P} \in \mathcal{R}^{2\times 2}$ such that

$$\begin{pmatrix} \bar{A}\bar{P}\bar{A} - \rho^2 \bar{P}^2 & \bar{A}\bar{P}\bar{B} \\ \bar{B}\bar{P}\bar{A} & \bar{B}\bar{P}\bar{B} \end{pmatrix} - \bar{X} \leq 0 \tag{22}$$

then set $P = \bar{P} \otimes I_p$ and define the Lyapunov function

$$V_k = \begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \end{pmatrix}^T P \begin{pmatrix} y_k - y_* \\ y_{k-1} - y_* \end{pmatrix} + f(y_k) - f(y_*), \tag{23}$$

which satisfies $V_{k+1} \leq \rho^2 V_k$ for all k. Moreover, we have $f(y_k) - f(y_*) \leq \rho^{2k} V_0$ for Nesterov's method.

$$\bar{P} = \begin{pmatrix} \sqrt{\frac{L}{2}} \\ \sqrt{\frac{m}{2}} - \sqrt{\frac{L}{2}} \end{pmatrix} \begin{pmatrix} \sqrt{\frac{L}{2}} & \sqrt{\frac{m}{2}} - \sqrt{\frac{L}{2}} \end{pmatrix}$$

clearly $\bar{P} \geq 0$. Now define $k = \frac{L}{m}$. Given $\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{k}-1}{\sqrt{k}+1}$ and $\rho^2 = 1 - \sqrt{\frac{m}{L}}$. It is easy to see that the left side of inequality (14) is as follows:

$$\frac{m(\sqrt{k}-1)^3}{2(k+\sqrt{k})} \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Continuous-time Dissipation Inequality

In systems and control theory one often encounters nonlinear control systems described, in the state space form, by means of a set of ordinary differential equations of the following type:

$$\begin{aligned} \dot{x} &= f(x) + G(x)u, \\ y &= h(x). \end{aligned} \tag{24}$$

Definition

The control system (24) is said to be dissipative with respect to the supply rate $S : \mathcal{R}^n \times \mathcal{R}^p \times \mathcal{R}^q \to \mathcal{R}$, if there exists a positive semidefinite storage function $V : \mathcal{R}^n \to \mathcal{R}$ such that the (integral) dissipation inequality

$$V(x(t_1)) - V(x(t_2)) \leq \int_{t_0}^{t_1} S(x(t), u(t), y(t))dt \tag{25}$$

If the storage function V is smooth then the integral dissipation inequality (25) can be rewritten as

$$\dot{V}(x(t)) \leq S(x(t), u(t), y(t)). \tag{26}$$

Consider

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{27}$$

Theorem

Let $X(t) \in \mathcal{R}^{(p+q)\times(p+q)}$ and $X^T(t) = X(t)$ for all t. Consider the quadratic supply rate

$$S(x(t), u(t), t) = \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^T X \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} \tag{28}$$

If there exists a family of matrices $0 \geq P(t) \in \mathfrak{R}^{p+q}$ such that

$$\begin{pmatrix} A^T P + PA + \dot{P} & PB \\ B^T P & 0 \end{pmatrix} - X \leq 0 \tag{29}$$

then the dissipation inequality (26) holds for all trajectories of (27) with $V(x, t) = x^T P x$.

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

Set

$$x = [\dot{Y}^T \; Y^T - y_*^T]^T, \quad u = \nabla f(y)$$

$$A(t) = \begin{pmatrix} \frac{-3}{t} I_p & 0_p \\ I_p & 0_p] \end{pmatrix}, \quad B(t) = \begin{pmatrix} -I_p \\ 0_p \end{pmatrix}$$

Denote

$$G(Y,t) = t^2(f(Y) - f_*)$$

Convexity implies

$$f(Y) - f_* \leq \nabla f(Y)^T (Y - y_*)$$

$$2t(f(Y) - f_*) \leq \begin{pmatrix} \dot{Y} \\ Y - y_* \\ u \end{pmatrix}^T \begin{pmatrix} 0_p & 0_p & 0_p \\ 0_p & 0_p & tI_p \\ 0_p & tI_p & 0_p \end{pmatrix} \begin{pmatrix} \dot{Y} \\ Y - y_* \\ u \end{pmatrix}$$

since

$$\dot{G}(Y,t) = 2t(f(Y) - f_*) + t^2 \nabla f(Y)^T \dot{Y}$$

then

$$\dot{G} \leq \begin{pmatrix} \dot{Y} \\ Y - y_* \\ u \end{pmatrix}^T \begin{pmatrix} 0_p & 0_p & \frac{t^2}{2} I_p \\ 0_p & 0_p & tI_p \\ \frac{t^2}{2} I_p & tI_p & 0_p \end{pmatrix} \begin{pmatrix} \dot{Y} \\ Y - y_* \\ u \end{pmatrix}$$

For supply rate S

$$\begin{pmatrix} 0_p & 0_p & \frac{t^2}{2} I_p \\ 0_p & 0_p & t I_p \\ \frac{t^2}{2} I_p & t I_p & 0_p \end{pmatrix}$$

Set

$$P = 2[\frac{t}{2} I_p \quad I_p]^T [\frac{t}{2} I_p \quad I_p], \quad V = x^T P x$$

Therefore

$$\dot{V} \leq S \leq -\dot{G}$$

Future work

$$\text{Minimize } f(x)$$
$$\text{S.t. } g_i(x) \leq 0,\ i = 1, 2, \ldots m$$

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

$$\dot{x} = -\nabla f(x) - \sum_{i=1}^{m} \lambda_i \nabla g_i(x)$$
$$\dot{\lambda}_i = P(\lambda, g_i),\ i = 1, 2, \ldots, m$$

Aujol, J. F., Dossal, C., & Rondepierre, A. (2019). Optimal convergence rates for Nesterov acceleration. SIAM Journal on Optimization, 29(4), 3131-3153.

Ebenbauer, C., Raff, T., & Allgower, F. (2009, June). Dissipation inequalities in systems theory: An introduction and recent results. In Invited lectures of the international congress on industrial and applied mathematics (Vol. 2007, pp. 23-42).

Fazlyab, M., Ribeiro, A., Morari, M., & Preciado, V. M. (2018). Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. SIAM Journal on Optimization, 28(3), 2654-2689.

Hu, B., & Lessard, L. (2017). Dissipativity theory for Nesterov's accelerated method. In International Conference on Machine Learning (1549-1557). PMLR.

Lessard, L., Recht, B., & Packard, A. (2016). Analysis and design of optimization algorithms via integral quadratic constraints. SIAM Journal on Optimization, 26(1), 57-95.

Kottenstette, N., & Antsaklis, P. J. (2010). Relationships between positive real, passive dissipative, & positive systems. In Proceedings of the 2010 American control conference (409-416). IEEE.

Wilson, A. C., Recht, B., & Jordan, M. I. A Lyapunov analysis of momentum methods in optimization (2016). arXiv preprint arXiv:1611.02635.

Thank you for your attention.